Essays on Contract Theory and Behavioral Economics

by

Daniel Gottlieb

B.A., IBMEC Business School (2001)
M.A., Fundação Getulio Vargas (2004)

Submitted to the Department of Economics
in partial fulfillment of the requirements for the degree of

Doctor of Philosophy in Economics

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

June 2009

Signature of Author .....
Department of Economics
June 2009

Certified by .....
Muhamet Yildiz
Professor of Economics
Thesis Supervisor

Certified by .....
Bengt Holmström
Paul A. Samuelson Professor of Economics
Thesis Supervisor

Accepted by .....
Esther Duflo
Abdul Latif Jameel Professor of Poverty Alleviation and Development Economics
Chairman, Departmental Committee on Graduate Studies

Essays on Contract Theory and Behavioral Economics

by

Daniel Gottlieb

Submitted to the Department of Economics
on June 2009, in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy in Economics

Abstract

This thesis is a collection of essays on contract theory and behavioral economics. Chapter 1 proposes a model of choice under risk based on imperfect memory and self-deception. The model assumes that people have preferences over their own attributes and can, to some extent, manipulate their memories. It leads to a non-expected utility representation and provides a unified explanation for several empirical regularities: non-linear probability weights, small-stakes risk aversion, regret and the competence hypothesis. It also leads to endowment and sunk cost effects. The model implies that behavior will converge to the one predicted by expected utility theory after a choice has been made a sufficiently large number of times.

Chapter 2 develops a model of competition with non-exclusive contracts in a market where consumers are time-inconsistent. Non-exclusivity creates a stark asymmetry between immediate-costs goods and immediate-rewards goods. In the former, non-exclusivity does not affect the equilibrium and, when consumers are sophisticated, the efficient allocation is achieved. When consumers are partially naive, the optimal sales tax may be either positive or negative and depends on parameters that are hard to estimate. In the case of immediate-rewards goods, however, the equilibrium features marginal-cost pricing and is always Pareto inefficient. Moreover, the optimal tax does not depend on the consumers' degree of naiveté and is a function of parameters that are easy to assess.

Chapter 3 is based on a joint work with Aloisio Araujo and Humberto Moreira. It considers a job-market signaling model where signals convey two pieces of information. The model is employed to study countersignalling (signals nonmonotonic in ability) and the GED exam. A result of the model is that countersignalling is more likely to occur in jobs that require a combination of skills that differs from the combination used in the schooling process. The model also produces testable implications consistent with evidence on the GED: (i) it signals both high cognitive and low noncognitive skills and (ii) it does not affect wages.

Chapter 4, which is also based on joint work with Aloisio Araujo and Humberto Moreira, characterizes incentive-compatibility in models where types are multidimensional and the single-crossing condition may not hold. This characterization is used to obtain the optimal contracts in multidimensional screening as well as the equilibria in multidimensional signaling models. Then, I determine the implications of signaling and screening models when the single-crossing condition is violated. I show that the unique robust prediction of signaling is the monotonicity of transfers in (costly) actions. Any function from the space of types to the space of actions and an increasing transfer schedule can be rationalized as an equilibrium profile of many signaling

models. Apart from the monotonicity of transfers in actions, I obtain an additional necessary and sufficient condition in the case of screening. In one-dimensional models, this condition states that the principal's profit as a function of the agent's type must grow at a higher rate under asymmetric information than under symmetric information.

Thesis Supervisor: Muhamet Yildiz
Title: Professor of Economics

Thesis Supervisor: Bengt Holmström
Title: Paul A. Samuelson Professor of Economics

Contents

To my parents, Marcel and Eliane.

Acknowledgements

Caminante, son tus huellas el camino, y nada más;
caminante, no hay camino, se hace camino al andar.
Al andar se hace camino, y al volver la vista atrás
se ve la senda que nunca se ha de volver a pisar.
– Antonio Machado

This thesis is the culmination of several years of learning, which were made possible by the support and encouragement of many people I am obliged to thank explicitly:

I begin by thanking my parents Eliane and Marcel. My mother, Eliane, has played a fundamental role in my academic career since long before I decided to pursue my graduate studies. My father, Marcel, has provided invaluable support and encouragement. My debt to him goes well beyond the quote above. My brothers and sisters, Michel, Felipe, Giovanna, and Debora, have always been there for me. Living away from them was one of the most challenging parts of my five years at MIT. I am deeply thankful to my grandmother Franca for her support through my graduate studies as well.

I also thank my advisors Muhamet Yildiz and Bengt Holmström. I am especially indebted to Muhamet for generously devoting an enormous amount of his time and for giving me detailed comments on this thesis. Bengt has profoundly shaped my views on contract theory and has given me invaluable advice and perspective. I am also extremely thankful to Drazen Prelec for comments and suggestions for the first chapter of this thesis and for guiding me throughout the job market experience.

I am grateful to my coauthors and former advisors at Fundação Getulio Vargas, Aloisio Araujo and Humberto Moreira, and to Jean Tirole. The contributions of Aloisio and Humberto are self-evident since chapters 3 and 4 of this thesis are based on our joint work. Jean provided insightful comments on chapters 1, 2, and 3 of this thesis. I consider myself fortunate for having had the opportunity to discuss this work with him.

I also would like to thank the following people for the useful comments and the discussions we have had over the course of my time at MIT: Daron Acemoglu, Abhijit Banerjee, Mathias Dewatripont, Glenn Ellison, Xavier Gabaix, and Jim Poterba.

Finally, I thank all the people who have made my years in graduate school a great experi-

ence. I am particularly thankful to Nicolás Arregui, Eduardo Azevedo, Igor Barenboim, Moshe Cohen, and my officemates: Arthur Campbell, Florian Ederer, Johannes Spinnewijn, and Jesse Edgerton. Moshe also provided extremely helpful feedback on Chapter 1 of this thesis and on my job market presentation. I look forward to interacting and working with many of these friends and colleagues in the future.

Last but not least, I would like to extend my deepest gratitude to Sabrina Najman for her constant support and encouragement.

Chapter 1

Imperfect Memory and Choice under Risk

1.1 Introduction

Choices with uncertain outcomes are an important part of a person's life. The outcomes often depend on the person's own attributes (e.g., skill, knowledge, or competence) and, therefore, affect the individual's self-views. Choices that turn out to be wrong typically lead to self-doubt, while choices that turn out to be right enhance the person's self-image. Hence, a person who cares about self-image has an incentive to manipulate recollections and beliefs. Indeed, there is sizeable psychological evidence that people value a positive self-image and manipulate their memories (see Section 1.1.1).

This chapter analyzes how the concern for self-image affects an individual's behavior under risk when memory is imperfect. The model is based on two basic premises: First, individuals have preferences over their own attributes; Second, they can (to some extent) affect what they will remember. Both assumptions are largely supported by evidence from the psychology literature. Apart from these two assumptions, individuals are assumed to behave as in standard economic models. Their preferences satisfy the axioms of expected utility theory. Furthermore, individuals follow Bayes' rule and, therefore, are aware of their memory imperfection. The model ties the concept of self-deception together with several deviations from standard expected utility theory, such as ambiguity aversion, non-linear probability weights, risk aversion over lotteries

with small stakes, regret aversion and the competence hypothesis. It also leads to endowment and sunk cost effects.

In its simplest version, the model consists of a two-period decision problem. In the first period, an individual observes the realization of a signal $\sigma \in \{H, L\}$, which is informative about her attributes. Then, she chooses the probability of remembering the realization of the signal by engaging in memory manipulation. In the second period, the individual applies Bayes' rule to her recollection of the signal. Because Bayes' rule implies that, on average, the individual's interpretation of her recollections are correct, self-deception does not change her (ex-ante) expected self-views. Hence, from an ex-ante point of view, memory manipulation is wasteful and, therefore, the agent would prefer not to observe the realization of the signal. Nevertheless, after observing the signal, the individual has an incentive to manipulate her memory in order to improve her self-image.

The model leads directly to preferences for avoiding information: people prefer not to acquire certain information if the expected benefit from making an informed decision is lower than the costs of self-deception. Because individuals anticipate these costs, they may prefer to make uninformed decisions if the objective value of information is sufficiently low. This result contrasts with Blackwell's celebrated theorem, which states that additional information can never be harmful. It is consistent, however, with the large psychology literature that connects self-deception and information avoidance. For example, people may avoid health exams, especially if the value of information is not high enough (e.g. the disease is not easily treatable) and if being diagnosed with the disease significantly affects the person's self-image. Individuals may also engage in "self-handicapping" strategies, such as under-preparing for an examination or getting too little sleep before physical exercise, in order to reduce the informational content of the signal. They may also display a "fear of competition" since outcomes from competitors are often informative about the person's own attributes.

When outcomes $\sigma \in \{H, L\}$ consist of monetary payments, the individual's expected utility can be represented by

$$w(q)\, u_H + [1 - w(q)]\, u_L,$$

where u_s is the decision-maker's utility in the state where s occurs and q is the probability of state $s = H$. The probability weight $w(q)$ is lower than the actual probability q when outcomes

lead to memory manipulation. Hence, these preferences provide a self-deception explanation for non-expected utility and ambiguity aversion.

As in other models that admit non-expected utility representations, the decision-maker may reject gambles with small but positive expected value. The agent may also exhibit a gap between the maximum willingness to pay for a good and the minimum compensation demanded for the same good (endowment effect). However, unlike other non-expected utility models, the departure from linear weights in my model is directly related to the decision-maker's self-perceived attributes. This departure is consistent with experimental evidence suggesting that deviations from expected utility theory are associated with the lotteries' being correlated with the decision-maker's skill or knowledge (c.f., Heath and Tversky, 1991, Josephs et al., 1992, Fox and Tversky, 1995, Goodie, 2003, and Goodie and Young, 2007).[1] In particular, the model provides a formalization of the (informal) theory of regret aversion based on self-perception proposed by Josephs et al. (1992). According to this theory, individuals with low self-image are more likely to make choices that minimize the possibility of regret. While different patterns may also be consistent with the model, it is able to predict the behavior described by Heath and Tversky (1991), according to which individuals prefer a knowledge-based lottery instead of a knowledge-independent lottery *with the same expected probability of winning* if and only if the individual believes that the probability of a positive outcome is high (competence hypothesis).[2] The model also allows the decision maker to reject small gambles without imposing unrealistic degrees of risk aversion over large gambles.

Two applications illustrate the theory. Successful trading usually requires certain skills or knowledge. At the very least, the agent must form expectations about how much each good is worth. In more complex markets, future prices of the goods must also be estimated. Thus, the outcome of the trade is informative about the person's skills or knowledge. Since decision-makers avoid information correlated with skills or knowledge, they will accept a trade only if the expected benefit from the trade exceeds a certain positive threshold. Therefore, self-deception leads to an endowment effect.

The second application considers the influence of sunk decisions on behavior. In several

[1] See Subsection 1.4.2 for a more detailed discussion.

[2] The model is also consistent with behavior that Eliaz and Spiegler (2006) have shown to be inconsistent with the Psychological Expected Utility model of Caplin and Leahy (2001).

contexts, revising one's decision usually involves admitting that a wrong decision was made and, therefore, it is often informative to the person about her own skills or knowledge. My model provides a self-deception explanation for the influence of sunk decisions on behavior that is consistent with arguments from the literature in psychology.

In a repeated setting in which the person observes a sequence of signals and engages in memory manipulation after each signal is realized, the attitude towards risk converges to the one implied by expected utility theory. This result is consistent with the arguments that people do not exhibit ambiguity aversion over events that have been observed several times and that experts are subject to much less bias than beginners (e.g. List, 2003, List and Haigh, 2005).

The structure of the chapter is as follows. Section 1.1.1 briefly reviews the psychological evidence on the memory and the related literature in economics. Section 1.2 introduces and discusses the general framework. In Section 1.3, I describe the implications for information acquisition. Section 1.4 considers lotteries over money and provides a representation result. In Section 1.5, I analyze a repeated version of the model. Section 1.6 presents the two applications of the model. Section 1.7 summarizes the main results and discusses possible extensions. The appendix relaxes some assumptions from the model and presents the proofs of the propositions in the text.[3]

1.1.1 Related Literature

An Overview of the Psychology Literature

> Ego-involvement, or its absence, makes a critical difference in human behavior. When a person reacts in a neutral, impersonal, routine atmosphere, his behavior is one thing. But when he is behaving personally, perhaps excitedly, seriously committed to a task, he behaves quite differently. In the first condition his ego is not engaged; in the second, it is. (Gordon W. Allport, 1943, pp. 459).

Psychologists have largely documented a human tendency to deny or misrepresent reality to oneself (i.e., engage in self-deception). In general, people consider themselves to be "smart,"

[3]Appendix A relaxes the additive separability assumption made in Section 1.4; Appendix B considers naive decision-makers, who are unaware of their memory imperfection; Appendix C considers models with any finite number of possible states, and Appendix D presents the proofs.

"knowledgeable," and "nice." Information conflicting with this image is usually ignored or denied. Greenwald (1980, pp. 605), for example, argued that "[o]ne of the best established recent findings in social psychology is that people perceive themselves readily as the origin of good effects and reluctantly as the origin of ill effects." Similarly, Gollwitzer, Earle, and Stephan (1982, pp. 702), claimed that the "asymmetrical attributions after success and failure" is a "firmly established finding."

People are also more likely to remember successes than failures (Korner, 1950). After choosing between two different options, they tend to recall the positive aspects of the chosen option and the negative aspects of the forgone option (Mather, Shafir, and Johnson, 2003). Relatedly, individuals overestimate their achievements and readily find evidence that they possess attributes which they believe to be correlated with success in personal or professional life (Kunda and Sanitioso, 1989; Quattrone and Tversky, 1984). Success is usually attributed to one's own ability and effort, whereas failure tends to be attributed to bad luck or other external variables (Gollwitzer, Earle, and Stephan, 1982, Zuckerman, 1979).[4] In group settings, where each individual's contribution cannot be unequivocally determined, people tend to attribute to themselves a larger share of the group's outcome after a success and a smaller share after a failure (Johnston, 1967).

Self-assessments and the memory are intrinsically connected. In his *Essay Concerning Human Understanding*, Locke (1690) identified the self with memory. Mill (1829, Vol. 2, pp. 174) argued that "[t]he phenomenon of Self and that of Memory are merely two sides of the same fact." Modern cognitive psychologists define the self as the "mental representation of oneself, including all that one knows about oneself" (Kihlstrom et al., 2002). Therefore, a model of self-views should devote considerable attention to memory.

In psychology, the memory is typically viewed as imperfect and manipulable. Rapaport (1961), for example, conceived "memory not as an ability to revive accurately impressions once obtained but as the integration of impressions into the whole personality and their revival *according to the needs of the whole personality.*" Allport (1943) believed that self-deception was a mechanism of ego defense and the maintenance of self-esteem. Hilgard (1949, pp. 374) argued

[4]Van den Steen (2004) presents a model of rational agents with differing priors that generates these biases. Harbaugh (2008) provides a career concerns explanation for prospect theory.

that "the need for self-deception arises because of a more fundamental need to maintain or to restore self-esteem. Anything belittling the self is to be avoided." Festinger (1957) suggested that individuals have a tendency to seek consistency among their cognitions (i.e., beliefs and opinions). He labeled the discomfort felt when one is presented with evidence that conflicts with one's beliefs and the resulting effort to distort those beliefs or opinions cognitive dissonance. In a review of the recent literature in social psychology, Sedikides, Green, and Pinter (2004, pp. 165) describe people as "striving for a positive self-definition or the avoidance of a negative self-definition (...) at the expense of accuracy and truthfulness." According to them, "[m]emory serves the function of shielding a positive self-definition from negativity."

There are several reasons why people may want to believe in things that are not true. First, there may be a hedonic value of positive self-views so that people simply like to think that they have these attributes.[5] Second, as argued by Compte and Postlewaite (2004), a person may benefit from having overconfident beliefs in situations where emotions affect performance. Third, manipulating one's own beliefs may facilitate the deception of others. Thus, holding an optimistic view of oneself may help convincing others of one's own value.[6] Fourth, there may be a motivational value of belief manipulation. As Benabou and Tirole (2002) and Weinberg (2006) argued, confidence in one's ability may help the person set more ambitious goals and persist in adverse situations.

This essay abstracts from the exact reason why people may value a positive self-image. The model developed here is based on the two basic ideas discussed above. First, individuals have preferences over their attributes. Second, they can affect what they will remember. The essay focuses on how memory manipulations affect the person's attitudes towards risk.

As the opening quote from Allport demonstrates, psychologists have long realized that self-deception may change a person's behavior. Festinger (1957, pp. 3), for example, argued that "[w]hen dissonance is present, in addition to trying to reduce it, the person will actively avoid situations and information which would likely increase the dissonance." More recently, Josephs et al. (1992, pp. 27) argued that "[r]isky decisions are potentially threatening to self-esteem

[5] For example, in Schelling's (1985) theory of the mind as a consuming organ, self-views have a hedonic value.

[6] As argued by Trivers (2000, pp. 115), "[b]eing unconscious of ongoing deception may more deeply hide the deception. Conscious deceivers will often be under the stress that accompanies attempted deception." This argument is modelled formally by Byrne and Kurland (2001) in an evolutionary game.

because the chosen alternative will occasionally yield a less desirable outcome than would some other alternative. When a less desirable outcome does occur, it can sometimes lead people to doubt their judgement and ability, especially when the decision is an important one."

This chapter shows that incorporating self-deception in a standard model of choice can lead to a unified theory of choice under risk that is consistent with economic phenomena such as ambiguity aversion, risk aversion over lotteries with small stakes, regret, and the competence hypothesis. It also leads to endowment and sunk cost effects.

An Overview of the Literature on Imperfect Memory

The economic literature on imperfect memory can be divided in two strands. The first assumes that decision makers are naive and act as if they have not forgotten anything (Mullainathan, 2002). The other strand assumes that decision makers are sophisticated, so that they draw Bayesian inferences given that they might have forgotten things. This essay follows the latter approach and considers the case of rational decision makers subject to imperfect recall.[7] As suggested by Piccione and Rubinstein (1997), the resulting game of imperfect recall is solved by the principle of "multiself consistency," whereby decisions made in different stages are viewed as being made by different incarnations of the decision maker.

Models of limited memory are a special case of imperfect memory. They were originally proposed by Robbins (1956) in the mathematical statistics literature. He suggested a decision rule for choosing between two lotteries with unknown distributions that was conditional on a finite number of outcomes (finite memory). In a series of papers, Cover and Hellman characterized optimal solutions to some finite memory problems.[8] More recently, economists have independently studied optimal decision making subject to limited memory. Dow (1991) considered the behavior of a consumer looking for the lowest price. Wilson (2003) studied how limited memory leads to certain biases in belief formation. Hirshleifer and Welch (2002) considered informational cascades generated by players who observe actions but not the information leading to such actions.

In a sequence of papers, Benabou and Tirole have used imperfect memory frameworks to

[7] Appendix B considers the case of naive decision makers.

[8] See Hellman and Cover (1973) for a review of the main results in this literature.

study questions from the psychology literature. Based on the assumption that agents recalled actions but not their motivations, they have proposed theories of personal rules and internal commitments (Benabou and Tirole, 2004), prosocial behavior (Benabou and Tirole, 2006b), and identity and taboos (Benabou and Tirole, 2006c). Using a model of self-deception, Benabou and Tirole (2002, 2006a) analyzed the provision of self-motivation and the formation of collective beliefs and ideologies.

The model of memory presented here is general enough to allow for an agnostic view of the behavior of the memory system. It encompasses both Benabou and Tirole's self-deception framework and a static version of the limited memory framework as special cases. This essay is also connected to the economic literature on cognitive dissonance (Akerlof and Dickens, 1982, Rabin, 1994). This literature assumes that agents derive utility from their beliefs and that they can, at some cost, choose their beliefs. Separately, Lowenstein (1987), Caplin and Leahy (2001 and 2004), and Kőszegi (2006) have studied models with anticipatory emotions.[9]

1.2 General Framework

The Decision Problem

The model examines a decision maker (DM) who has preferences over her attributes θ. Attributes θ may be interpreted as skills, knowledge, or competence as well as a parameter of anticipatory utility. Let Θ be a non-empty subset of $\mathbb{R}$ representing the possible values of θ and let $F(.)$ denote the agent's prior distribution of θ.[10]

The DM acts in 3 periods ($t = 0, 1, 2$). In period 0, she chooses an action a from a non-empty, compact subset of a finite dimensional Euclidean space A. For example, a can be an investment decision or a decision of whether to undertake some medical examination. The set

[9] Brunnermeier and Parker (2005) proposed a theory of "optimal expectations," according to which individuals choose their beliefs balancing the gains from anticipating a higher future utility with the losses from suboptimal decision-making. Similarly, Hvide (2002) proposed the notion of "pragmatic beliefs," which are the beliefs that maximize the individual's utility. Bernheim and Thomadsen (2005) showed that memory imperfections and anticipatory emotions may lead to a resolution of Newcomb's Paradox and sustain cooperation in the Prisoners Dilemma.

[10] Θ can be continuous or discrete, as long as it contains at least two elements (otherwise, θ cannot be random). Note that we have not assumed that the agent has a correct prior distribution over θ. Therefore, agents are allowed to hold optimistic or pessimistic beliefs about their attributes.

A can also be a singleton, in which case the agent makes no choice in period 0.[11]

In period $t = 1$, an outcome σ_a, which can be either high (H) or low (L), is observed. The outcome σ_a may be a purely informative signal, entering the agent's preferences only indirectly through her beliefs about her attributes θ. It may also affect the agent's preferences directly. For example, a medical exam consists of a purely informative signal, whereas the outcomes of an investment affect an individual not only through their informational content but also through the different monetary payments associated with them. I denote by $q_a \in (0,1)$ the probability of observing a high outcome given action $a \in A$. A high outcome is assumed to be more favorable than a low outcome in the sense of first-order stochastic dominance:

$$F(\theta|\sigma_a = H) \leq F(\theta|\sigma_a = L) \text{ for all } \theta \in \Theta, \tag{1.1}$$

with strict inequality for some value of θ, and for all $a \in A$.

Following Rabin (1994), Benabou and Tirole (2002, 2006a), and Benabou (2008), I assume that the individual can, at a cost, influence her recollections. The DM remembers the outcome $s \in \{H, L\}$ with probability

$$\eta_s + m_s,$$

where the parameter $\eta_s \in [0,1]$ is the agent's "natural" rate of remembering outcome s. This rate determines the probability that the DM recollects the outcome if she does not employ any manipulation effort. However, the DM is also able to depart from the natural rate of forgetting the outcome by exerting effort $m_s \in [-\eta_s, 1-\eta_s]$ in period $t = 1$. Engaging in memory manipulation m_s leads to a cost of $\psi_s(m_s) \geq 0$, $s \in \{H, L\}$. The agent's recollection of the outcome σ_a is denoted by $\hat{\sigma}_a \in \{H, L, \varnothing\}$, where we write $\hat{\sigma}_a = \varnothing$ if the outcome has been forgotten.

In period $t = 2$, the DM takes an action b in a non-empty, compact subset of a finite dimensional Euclidean space B. For example, b can be a decision of whether to continue with some previous investment or whether to undertake some medical treatment. B can also be a singleton, in which case the DM does not act after observing the outcome. Figure 1-1 presents the informational structure.

[11] In some applications, the set A may also include the possibility of not observing a signal.

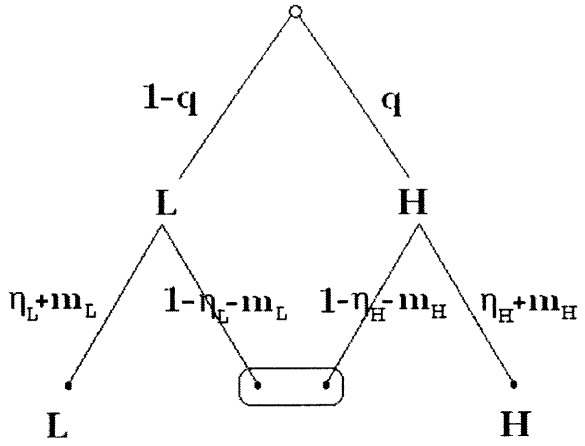

Figure 1-1: Informational Structure

Preferences satisfy the standard axioms of expected utility theory. Therefore, there exists utility function $u : \Theta \times A \times B \times \{H, L\} \to \mathbb{R}$ representing the DM's preferences. Furthermore, $u(\theta, a, b, \sigma)$ is strictly increasing in θ for all $(a, b, \sigma) \in A \times B \times \{H, L\}$.

When $u(\theta, a, b, H) = u(\theta, a, b, L)$ for all $(\theta, a, b) \in \Theta \times A \times B$, we refer to outcomes as signals since they do not affect the agent's utility directly. In this case, we say that the model has *purely informative signals.* When signals are purely informative and A and B are singletons, we say that signals have a *purely hedonic value.* In models where signals have a purely hedonic value, the DM does not need to take any decision and the only reason for memory manipulation is the improvement of the individual's self-views.

We refer to the case where $u(\theta, a, b, H) > u(\theta, a, b, L)$ for all $(\theta, a, b) \in \Theta \times A \times B$ as a model of *monetary outcomes.* In this case, outcomes are interpreted as monetary payments and a high outcome raises the agent's utility both directly and through beliefs about θ.[12]

The cost of memory manipulation ψ_s can be related to psychic costs (stress from repression of negative information or effort to focus on positive information), time (searching for reassuring information or excuses, lingering over positive feedback), or real resources (avoiding certain cues and interactions or eliminating evidence). They can also be interpreted as the shadow costs

[12] Although the case described above, where the outcome with a higher monetary payment provides more favorable news about the DM's attributes, is the most intuitive, this is not necessary for our results. Alternatively, one could assume that the outcome with a higher monetary payment is bad news about the DM's attributes.

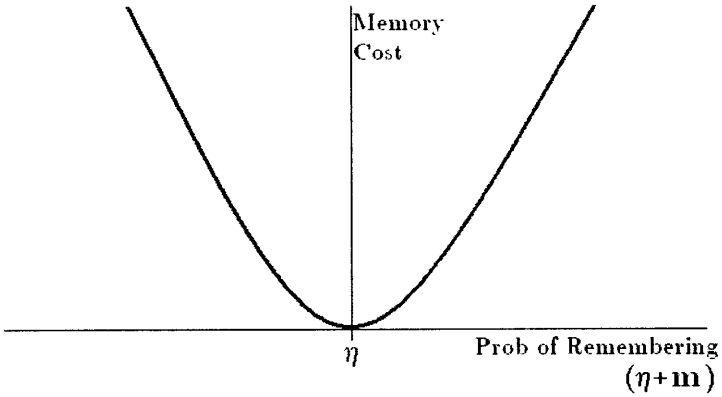

Figure 1-2: Cost of Memory

of memory in a limited information framework. Remembering an outcome with probability above its natural rate η_s requires an individual to focus on it and on information correlated with it. In turn, this restricts the amount of attention available to other information (which has shadow cost ψ_s). Similarly, forgetting an outcome with probability above the natural rate $1-\eta_s$ requires an individual to focus on confronting evidence which again restricts the amount of attention available to other potentially useful information.[13]

Assumption 1 The cost of memory manipulation $\psi_s(m_s)$ is strictly decreasing in $m_s < 0$, strictly increasing in $m_s > 0$, convex, twice-continuously differentiable, and such that $\psi_s(0) = 0$, $s \in \{H, L\}$.

Figure 1-2 depicts the costs of memory manipulation implied by Assumption 1. I further assume that the agent forgets a high outcome with some positive probability if she does not exert any effort:[14]

Assumption 2 $\eta_H < 1$.

[13]For example, Steele's (1988) self-affirmation theory argues that people cope with negative outcomes in one domain by focusing in other, unrelated domains.

[14]If $\eta_H = 1$, then the model becomes trivial. Since the agent always recalls high outcomes, she will perfectly infer that $\sigma = L$ was observed if she recollects $\hat{\sigma} = \varnothing$. Therefore, she will never engage in memory manipulation.

The model can also be seen as a conflict between a "hot" or "impulsive" self and a "cold" self. The hot self (self 1) wants to minimize current losses from negative information and maximize the current gains from positive information.[15] The cold self (self 2) wants to circumvent the manipulations made by the hot self in order to make a correct inference. The hot self exerts efforts m_L and m_H in order to manipulate the beliefs of the cold self. Then, the cold self applies Bayes' rule in order to filter these manipulations and make a correct decision b.[16]

As the following examples show, the general framework encompasses other models of imperfect memory.

Example 1 (The Forgetfulness Model of Benabou and Tirole, 2002) *Take* $\eta_L = 1, \eta_H = 0$ *and* $\psi_H(m_H) = +\infty$ *for all* $m_H > 0$ *so that high outcomes are always forgotten (i.e.,* $\eta_H + m_H = 0$*). Figure 1-3 presents the informational structure in this case. This is the memory framework from Benabou and Tirole (2002). It can be interpreted as a model of bad news or no news. If the agent receives bad news, she can exert an effort* $m_L \in [-1, 0]$ *in order to forget them.*

If the state $\varnothing$ is reinterpreted as the recollection of a high outcome, then the model from Example 1 becomes one where the agent is able to convince herself that a low outcome was a high outcome.[17] Hence, memory manipulation would allow the DM to believe that she observed an outcome $\sigma = H$. This reinterpretation is compatible with neurological evidence from Prelec (2008), who showed that subjects experience heavy brain activity only when they try to convince themselves that a bad outcome was actually a good one. In the other states (both when they acknowledge a mistake or when they believe to have been correct), no such activity is detected. Hence, Example 1 can be interpreted as the agent incurring psychological costs when she tries

[15] This interpretation assumes that the hot self is rational in the sense of taking into account the benefits and costs of memory manipulation. Several papers in social psychology have documented that individuals tend to be more realistic and impartial when making important decisions (c.f., Taylor and Gollwitzer, 1995, and references therein). Therefore, self-deception seems to decrease when the cost of a mistake increases. Prelec (2008) presented experimental evidence where self-deception responds positively to its expected benefits.

Similarly to this interpretation, Bodner and Prelec (2002) present a signaling model between an agent's privately informed gut and the agent's uninformed mind.

[16] The model can be interpreted as a formalization of the neurophysiological argument put forth by Trivers (2000). According to this interpretation, self 1 would be the person's unconscious process of information manipulation. In the context of intertemporal choice, several papers have proposed dual self models (c.f. Thaler and Shefrin, 1981, Fudenberg and Levine, 2006, and Brocas and Carrillo, 2008).

[17] In this model, the agent would never choose to believe that a high outcome was actually low.

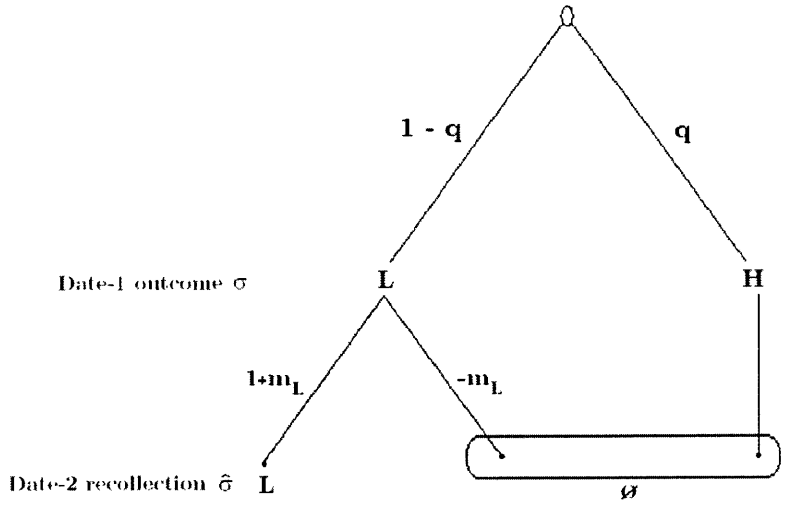

Figure 1-3: Forgetfulness Model

to convince herself that a bad outcome was actually a good one.

Example 2 (The Limited Memory Model) *Take* $\eta_L = \eta_H = 0$ *so that the DM forgets any outcome if she does not employ memory efforts. Then, the framework becomes a model of limited memory. In this model, the DM must allocate a limited amount of memory in order to store information. By spending a memory cost* $\psi_s(m_s)$, *she remembers an outcome* $s \in \{H, L\}$ *with probability* m_s. *A higher effort* m_s *can be interpreted as having greater memory resources used to store the information.*[18]

The following examples present applications of the general framework to specific environments:

Entrepreneurship Example An employed individual is considering quitting her job and starting a new company. Building a successful company requires certain entrepreneurial skills which are unknown to the individual. Therefore, a success provides favorable news about the individual's skills. If she decides not to quit her job, the individual obtains a wage $w \in \mathbb{R}_+$ and does not learn any information about her skills.

[18] *Dow (1991) considers a consumer who searches sequentially for the lowest price, but who only remembers each price as belonging to a finite number of categories. Wilson (2003) considers a decision-maker who must act after a large number of periods but whose memory is restricted to a finite number of states.*

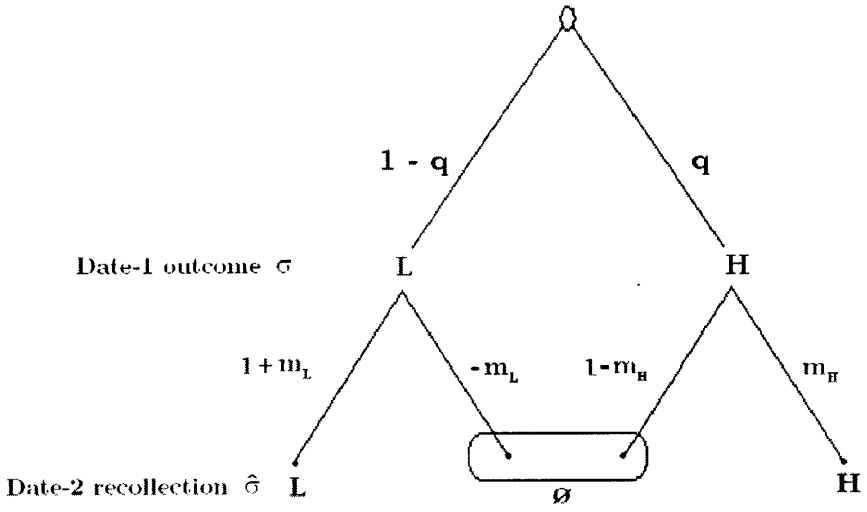

Figure 1-4: Limited Memory Model

In this essay, this situation is modeled as follows. Let the individual's career choice be denoted by $a = E$ if she becomes an entrepreneur and by $a = W$ if she remains a worker and let θ denote the individual's entrepreneurial skills. The outcome from starting a company is denoted by σ, which is equal to H in the case of success and L in the case of failure. After the outcome σ is observed, the entrepreneur may engage in memory manipulation. In this model, there is no ex-post choice (B is a singleton). The agent's decision tree is presented in Figure 1-5.

Appendix C considers a more general model. In that model, an outcome is a vector $\sigma = (s, r)$ consisting of a binary variable reflecting whether or not the company was successful, $s \in \{S, F\}$, and an external variable $r \in \mathbb{R}$ which affects the outcomes but is independent of the agent's attributes (e.g., general market conditions, economy-wide shocks). The entrepreneur always remembers whether the company succeeded or failed but may forget the prevailing external conditions r.

Succeeding under adverse conditions provides good news about the individual's skills. Similarly, failing under favorable conditions is bad news about her skills. In this model, that the agent will manipulate her memory in order to forget positive external shocks and remember negative shocks. This result is consistent with the psychological literature described in Section

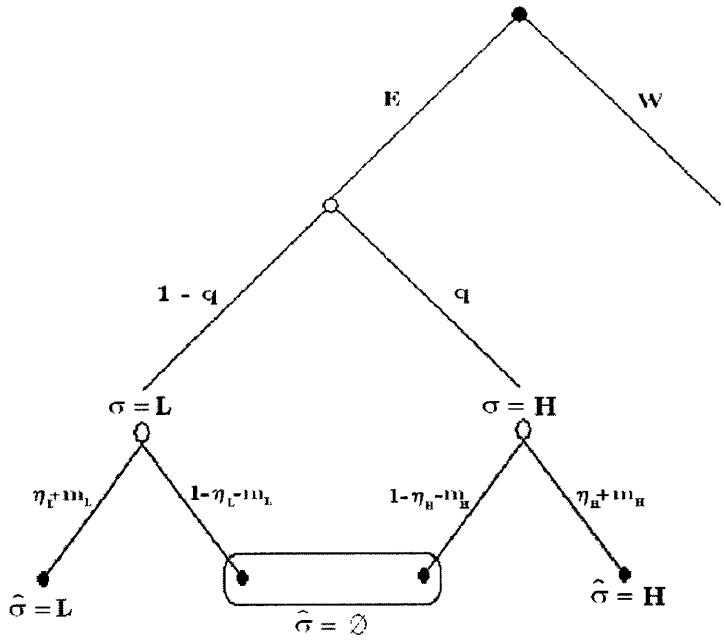

Figure 1-5: Entrepreneurship Example

1.1.1, which shows that success is usually attributed to one's own attributes whereas failure tends to be attributed to bad luck or other external variables.

In Section 1.4, I will show that self-deception will prevent some individuals from becoming entrepreneurs even when the expected monetary payoffs from starting a new company are higher than the payoff from remaining on the previous job.

Used Car Example An individual is considering whether to purchase a used car or to use public transportation. A used car may be defective. Moreover, detecting whether the car is defective requires certain skills. Therefore, purchasing a defective car conveys unfavorable information about the buyer's skills and requires the car to be fixed. If she decides to use public transportation, no information is learned.

This situation is modeled as follows. Let $a = C$ denote the choice of purchasing a used car and let $a = PT$ denote the choice of using public transportation. Denote by $\sigma = H$ the case where the car is non-defective and $\sigma = L$ the case where it is defective. After the consumer learns that the car was defective, she may manipulate her memory in order to forget that it needed to be fixed. Similarly, if the car was non-defective, she may exert some effort

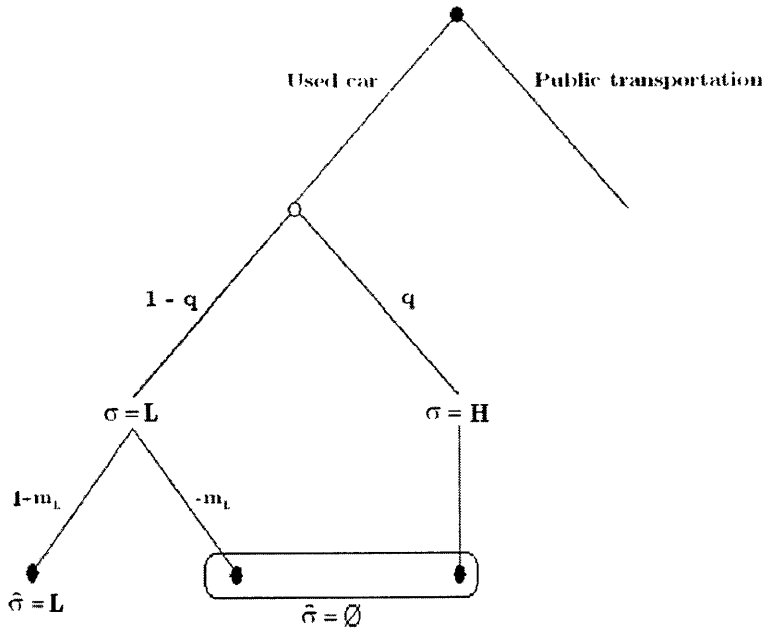

Figure 1-6: Used Car Example

to remember that the car did not need to be fixed. Assuming the memory system from the forgetfulness model of Example 1, we obtain the decision tree depicted in Figure 1-6.

Section 1.4 will show that if the expected monetary benefit from buying the used car is positive but lower than the expected self-deception costs, the individual will prefer not to purchase it.

1.2.1 Modeling as a Multiself Game

This essay follows Piccione and Rubinstein (1997) in modeling a decision problem with imperfect memory as a game between different selves. The decision maker is treated as a collection of selves, each of them unable to control the behavior of future selves. As will be described in Subsection 1.2.2, the decision made by an agent with imperfect recall corresponds to the perfect Bayesian equilibrium (PBE) of this game between selves.[19]

The extensive form of the multiself game is presented in Figure 1-7. There are two players: self 1 and self 2. Both selves have the same utility functions but different information sets. In

[19]For the games considered here, the set of sequential equilibria coincides with the set of PBE.

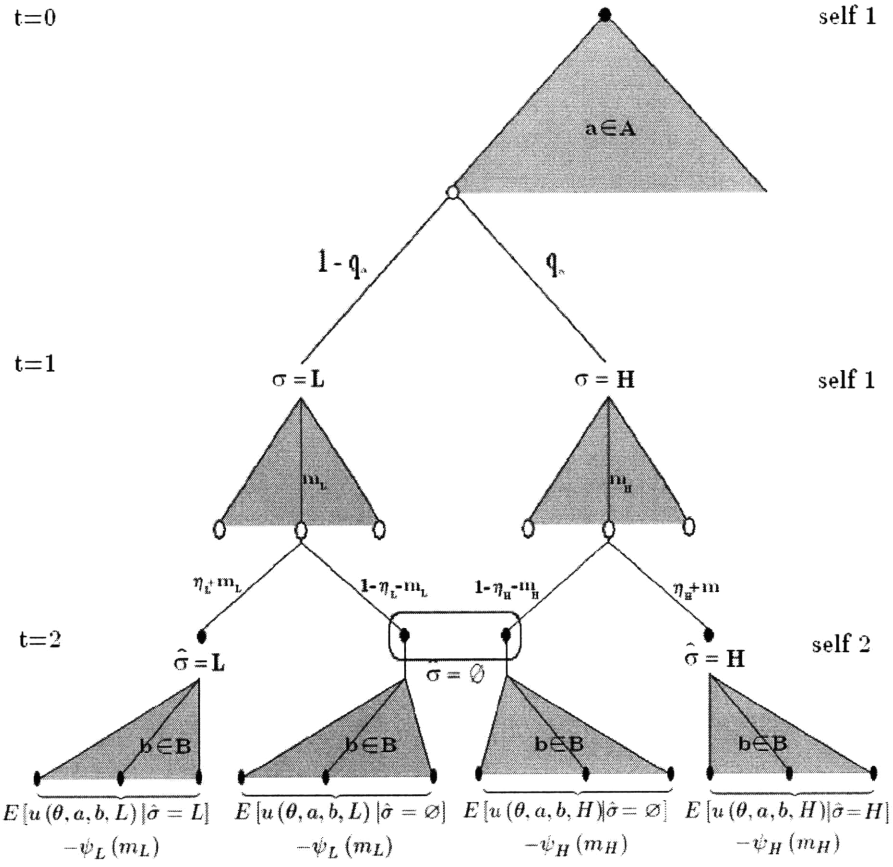

Figure 1-7: Extensive Form

period 0, self 1 chooses an action $a \in A$. Then, nature plays a high outcome with probability q_a and a low outcome with probability $1 - q_a$. In period 1, conditional on the outcome $s \in \{H, L\}$, self 1 decides the amount of memory manipulation m_s. Then, given the outcome s and the manipulation effort m_s, nature plays $\hat{\sigma} = s$ with probability $\eta_s + m_s$ and $\hat{\sigma} = \varnothing$ with probability $1 - \eta_s - m_s$. In period 2, self 2 observes the recollection $\hat{\sigma}$ and takes an action $b \in B$. Then, both selves get payoff $E[u(\theta, a, b, \sigma) | \hat{\sigma}] - \psi_s(m_s)$.

Because the DM has preferences over θ, she has an interim incentive to manipulate her beliefs by exerting effort m_s. However, the set of possible beliefs that an agent can hold is restricted by the assumption that recollections are interpreted according to Bayes' rule. Thus, the agent makes correct inferences about her attributes θ given her recollections $\hat{\sigma}$.

Equivalently, we can conceptualize an "inferential self" who tries to make a correct inference about the agent's attributes given the recollections. This inferential self chooses the agent's expected utility so as to minimize a quadratic loss function:

$$u_{\hat{\sigma}}(a,b,\sigma) = \arg\min_{\hat{u}\in\mathbb{R}} \int_{-\infty}^{\infty} [\hat{u} - u(\theta,a,b,\sigma)]^2 \, dF(\theta|\hat{\sigma}).$$

The solution to this program is $u_{\hat{\sigma}}(a,b,\sigma) = \int u(\theta,a,b,\sigma)\, dF(\theta|\hat{\sigma})$, which is the Bayes estimator of $u(\theta,a,b,\sigma)$ given the recollection $\hat{\sigma}$. Thus, by minimizing a quadratic loss function, the inferential self constrains the decision-maker to be a Bayesian given her memory imperfection.

Remark 1 *Denote the expected value of θ conditional on the observed outcome σ_a by θ_{σ_a} and the expected attributes conditional on the recollection $\hat{\sigma}_a$ by $\hat{\theta}_{\hat{\sigma}_a}$. $\hat{\theta}_{\hat{\sigma}_a}$ is "less variable" than θ_{σ_a} in the sense of second-order stochastic dominance.*[20] *Therefore, because θ_{σ_a} is the Bayes estimate of θ given the outcome σ_a, forgetfulness implies that the decision-maker updates observed outcomes σ_a less than implied by Bayes' rule. This result is consistent with experimental evidence from Falk, Huffman, and Sunde (2006).*

1.2.2 Solution Concept

As described in the previous subsection, the decision made by an agent with imperfect recall is modeled as the perfect Bayesian equilibrium (PBE) of the multiself game. Let $\mu(.|\hat{\sigma})$ denote the DM's posterior beliefs about θ given $\hat{\sigma}$ and let $E_\mu[.|\hat{\sigma}]$ denote the expectation operator with respect to $\mu(.|\hat{\sigma})$. Given a profile of memory manipulation manipulation (m_L, m_H), let $E_{\hat{\sigma}_a}[.|m_L, m_H]$ denote the expectation with respect to the distribution of $\hat{\sigma}_a$.

Definition 1 *A PBE of the game is a strategy profile $(a^*, b^*, m_H^*(a), m_L^*(a))$ and posterior beliefs $\mu(.|\hat{\sigma}_a)$ such that:*

1. $$a^* \in \arg\max_{a\in A} \left\{ \begin{array}{c} E_{\hat{\sigma}_a}\left[E_\mu\left[u\left(a, b_a^*(\hat{\sigma}_a), \theta, \sigma_a\right)|\hat{\sigma}_a\right]|m_L^*(a), m_H^*(a)\right] \\ -q\psi_H(m_H^*(a)) - (1-q)\psi_L(m_L^*(a)) \end{array} \right\};$$

2. $$m_s^*(a) \in \arg\max_{m_s} \left\{ \begin{array}{c} (\eta_s + m_s) E_\mu[u(a, b_a^*(\hat{\sigma}_a), \theta, s)|\hat{\sigma}_a = s] \\ +(1-\eta_s - m_s) E_\mu[u(a, b_a^*(\hat{\sigma}_a), \theta, s)|\hat{\sigma}_a = \varnothing] - \psi_s(m_s) \end{array} \right\},$$

[20] *See Appendix D for the proof.*

$s \in \{H, L\}$;

3. $b_a^*(\hat{\sigma}) \in \arg\max_{b \in B} \{E_\mu [u(a, b, \theta, \sigma_a) | \hat{\sigma}_a = \hat{\sigma}]\}$;

4. $\mu(\theta|\hat{\sigma}_a = \hat{\sigma})$ *is obtained by Bayes' rule if* $\Pr(\hat{\sigma}_a = \hat{\sigma}|m_L^*(a^*), m_H^*(a^*)) > 0, \forall \hat{\sigma} \in \{L, H, \varnothing\}$.

Conditions $1-3$ are the standard sequential rationality conditions. Condition 1 states that self 1 chooses an ex-ante action a that maximizes the agent's expected utility in period 0 given the behavior of self 2. Condition 2 states that, conditional on each outcome $s \in \{H, L\}$, self 1 chooses the amount of manipulation that maximizes the her expected payoff. Condition 3 states that self 2 takes an action that maximizes her utility given the beliefs she holds about the manipulation employed by self 1.

Condition 4 is the standard consistency condition, requiring that beliefs of self 2 satisfy Bayes' rule given the strategy of self 1. For every recollection $\hat{\sigma}$ that is reached with positive probability, it implies that $\mu(\theta|\hat{\sigma}_a) = F(\theta|\hat{\sigma}_a)$. Because of Bayesian updating, Condition 3 becomes

$$b_a^*(\hat{\sigma}) \in \arg\max_b \int u(a, b, \theta, \sigma) \, dF(\theta|\hat{\sigma}_a = \hat{\sigma}),$$

for any recollection $\hat{\sigma}$ that is reached with positive probability. The following proposition establishes the existence of a PBE:

Proposition 1 (Existence) *There exists a PBE.*

Define the expected utilities given $\sigma_a = H$ and $\sigma_a = L$ by

$$\begin{aligned} u_H(a, b, \sigma_a) &\equiv \int u(a, b, \theta, \sigma_a) \, dF(\theta|\sigma_a = H), \text{ and} \\ u_L(a, b, \sigma_a) &\equiv \int u(a, b, \theta, \sigma_a) \, dF(\theta|\sigma_a = L). \end{aligned} \tag{1.2}$$

Given the recollection of a high signal, $\hat{\sigma}_a = H$, self 2 infers that a high signal was observed in period 1. Hence, Bayesian updating implies that the expected utility of self 1 conditional on $\hat{\sigma}_a = H$ is $u_H(a, b_a(H), H)$. Similarly, the expected utility of self 1 conditional on $\hat{\sigma}_a = L$ is $u_L(a, b(L), L)$.

Let $m_L^*(a)$ and $m_H^*(a)$ denote the amount of memory manipulation that self 2 believes was employed in period 1. Note that the PBE concept implies that $m_L^*(a)$ and $m_H^*(a)$ are taken as

given by self 1 when choosing the amount of memory manipulation to exert. If the DM forgets which signal was observed in period 1 (i.e., she recollects $\hat{\sigma}_a = \varnothing$), then there is a probability $(1-q_a)(1-\eta_L - m_L^*(a))$ that $\sigma_a = L$ was observed and a probability $q_a(1-\eta_H - m_H^*(a))$ that $\sigma_a = H$ was observed. Thus, the expected utility given $\hat{\sigma}_a = \varnothing$ is

$$\begin{aligned} u_\varnothing(a, b_a(\varnothing), \sigma_a) \equiv\ & \alpha(m_L^*, m_H^*)\, u_H(a, b_a(\varnothing), \sigma_a) \\ & + [1 - \alpha(m_L^*, m_H^*)]\, u_L(a, b_a(\varnothing), \sigma_a), \end{aligned} \tag{1.3}$$

where $\alpha(m_L, m_H) \equiv \frac{q_a(1-\eta_H - m_H)}{q_a(1-\eta_H - m_H)+(1-q_a)(1-\eta_L - m_L)}$ is the conditional probability of $\sigma_a = H$ implied by Bayes' rule.

Conditions 2 and 3 from Definition 1 state that, after observing signal $\sigma_a = s \in \{H, L\}$, self 1 chooses m_s to maximize

$$(\eta_s + m_s)\, u_s(a, b_a(s), s) + (1 - \eta_s - m_s)\, u_\varnothing(a, b_a(\varnothing), s) - \psi_s(m_s).$$

Using equation (1.3), the expected utility after a low signal can be written as

$$\begin{aligned} & u_L(a, b_a(\varnothing), L) + (\eta_L + m_L)[u_L(a, b_a(L), L) - u_L(a, b_a(\varnothing), L)] \\ & + (1 - \eta_L - m_L)\, \alpha(m_L^*(a), m_H^*(a))[u_H(a, b_a(\varnothing), L) - u_L(a, b_a(\varnothing), L)] - \psi_L(m_L) \end{aligned} \tag{1.4}$$

Note that self 1 takes three factors into account when choosing the amount of effort to forget bad news. First, forgetting a low signal leads to a higher utility through a more favorable inference about θ since $u_H(a, b_a(\varnothing), L) > u_L(a, b_a(\varnothing), L)$ (self-deception factor). Second, it leads to a sub-optimal choice of b since $u_L(a, b_a(L), L) \geq u_L(a, b_a(\varnothing), L)$ (decision-making factor). Third, self-deception leads to a memory cost of $\psi_L(m_L)$ (memory cost factor).

Analogously, conditional on a high signal, self 1 chooses m_H to maximize:

$$\begin{aligned} & (\eta_H + m_H) \left\{ \begin{array}{c} [1 - \alpha(m_L^*(a), m_H^*(a))] \\ + u_H(a, b_a(H), H) - u_H(a, b_a(\varnothing), H) \end{array} \right\} \\ & + u_\varnothing(a, b_a(\varnothing), H) - \psi_H(m_H). \end{aligned} \tag{1.5}$$

This equation displays the three factors that determine the amount of effort to remember good news. First, remembering a high signal leads to a higher utility through a more favorable inference about θ since $u_H(a, b_a(\varnothing), H) > u_L(a, b_a(\varnothing), H)$. It also leads to better decision-making since $u_H(a, b_a(H), H) > u_H(a, b_a(\varnothing), H)$. However, it leads to a memory cost of $\psi_H(m_H)$.

The improvement in decision-making leads the DM to engage in an effort to remember a high signal. The effect from self-image also leads the DM to exert an effort to remember the high signal. Because small amounts of memory manipulation have second-order costs, the DM always remembers a high signal with probability above her natural rate η_H :

Proposition 2 (Remembering Good News) *Suppose that ψ_s is strictly convex, $s \in \{H, L\}$. Then, in any PBE, $m_H^*(a) > 0 \ \forall a \in A$.*

The DM's ex-ante expected utility (in period 0) is

$$E_{\hat{\sigma}_a}\left[E_\mu\left[u\left(\theta, a, b_{a^*}(\hat{\sigma}), \sigma\right)|\hat{\sigma}\right]\right] - q\psi_H\left(m_H^*(a^*)\right) - (1-q)\,\psi_L\left(m_L^*(a^*)\right). \tag{1.6}$$

As in other decision problems with imperfect recall, the timing of decisions has important implications for the solution. If the agent could commit to a strategy at an ex-ante stage, she would generally choose a different amount of memory manipulation.

Consider, for example, the model of purely hedonic signals. In this case, equations (1.4) and (1.5) imply that the DM faces a trade-off between self-deception and memory costs. Manipulating one's memory into forgetting a low signal directly increases the individual's expected payoff by raising the probability that the signal is forgotten. Similarly, exerting effort to remember a high signal directly raises her expected payoff by decreasing the probability that the signal is forgotten. However, these manipulations also decrease the DM's expected payoff indirectly by reducing the relative probability of a high signal when the signal is forgotten. Bayesian updating implies that the indirect effects exactly cancel the direct effects out. Because the DM is not fooled on average, she adjusts the expected attributes given $\hat{\sigma} = \varnothing$ to take into account the relative frequency that each signal is forgotten. Therefore, from an ex-ante perspective, memory manipulation only leads to memory costs and the DM would prefer not to engage in memory manipulation at all ($m_H = m_L = 0$). However, the multiself approach implies that self

1 does not take into account the indirect effects from memory manipulation and, therefore, chooses to engage in memory manipulation.[21] Hence, unlike in decision problems with perfect recall where ex-ante optimal strategies are always time-consistent, the ex-ante optimal strategy is time-inconsistent.[22]

In cases where outcomes affect ex-post actions b (i.e., information has positive value), it is ex-ante optimal to choose some positive amount of memory manipulation.[23] In these cases, the optimal strategy from an ex-ante perspective would always have a probability to remember (weakly) above the natural rate η_s, $s \in \{L, H\}$.

Recall that self 1 takes three factors into account when choosing the amount of memory manipulation: (i) self-deception, (ii) decision-making, and (iii) memory costs. As discussed previously, Bayesian updating implies that the self-deception effect vanishes from the DM's ex-ante utility. Since only factors (ii) and (iii) would be taken into account, the DM would choose to remember good news and to forget bad news less frequently if she could commit to a strategy in period 0. Let the ex-ante expected utility be denoted by

$$\begin{aligned} \mathcal{U}\left(m_H, m_L, a, \{b(\hat{\sigma})\}_{\hat{\sigma}\in\{H,L,\varnothing\}}\right) &= E_{\hat{\sigma}_a}\left[E_\mu\left[u(\theta, a, b(\hat{\sigma}), \sigma)\,|\hat{\sigma}\right]\right] \\ &\quad -q\psi_H(m_H) - (1-q)\,\psi_L(m_L). \end{aligned}$$

Proposition 3 establishes this claim formally:

Proposition 3 (Excessive Manipulation) *Let* $\left(\tilde{m}_H(a), \tilde{m}_L(a), \left\{\tilde{b}_a(\hat{\sigma})\right\}_{\hat{\sigma}\in\{H,L,\varnothing\}}\right)$ *be a maximizer of* $\mathcal{U}$ *given action* a *and suppose* $\mathcal{U}$ *is a concave function of* m_H *and* m_L.[24] *Then, in any PBE with manipulations* $m_H^*(a)$ *and* $m_L^*(a)$,

$$m_H^*(a) \geq \tilde{m}_H(a) \qquad \text{and} \qquad m_L^*(a) \leq \tilde{m}_L(a)$$

[21] Note that the DM would never choose to undo the memory manipulation in period $t = 2$ and find out the true outcome σ if she had a chance to do so.

[22] See Piccione and Rubinstein (1997) for a discussion of decision problems with imperfect recall. In the present model, because all nodes are reached with positive probability, the two equilibrium concepts proposed there (multiself consistency and modified multiself consistency) coincide.

[23] More precisely, let $b(\hat{\sigma})|_{m_L,m_H}$ denote the action that maximizes the DM's utility given recollection $\hat{\sigma}$ and conditional on manipulation efforts m_L and m_H. Then, $b(H)|_{m_L=m_H=0} \neq b(\varnothing)|_{m_L=m_H=0}$ implies that the manipulation effort m_H that maximizes the ex-ante expected utility is strictly positive. Analogously, if $b(L)|_{m_L=m_H=0} \neq b(\varnothing)|_{m_L=m_H=0}$ then m_L that maximizes the ex-ante expected utility is strictly positive.

[24] *It is straightforward to show that* $\mathcal{U}$ *is always a concave function of* m_H *and* m_L *when* B *is a singleton.*

for all $a \in A$, with at least one of the inequalities being strict.

1.2.3 Equilibrium when Information has Purely Hedonic Value

In order to illustrate the impact of self-deception on choice, this subsection considers the simple case where signals are purely informative and the DM does not take any action (i.e., information has purely hedonic value). In this case, the only reason for memory manipulation is the improvement of the DM's self-views. Since remembering a low signal decreases self 1's expected utility, she would never choose manipulate her memory in order to remember a low signal (i.e., $m_L^* \leq 0$). Analogously, she would never manipulate her memory so as to forget a high signal (i.e., $m_H^* \geq 0$).

Since, in the purely hedonic case considered in this subsection, A and B are singletons and the outcome of the signal does not enter the agent's utility directly, I omit the terms a, b, and σ_a from the DM's von Neumann-Morgenstern utility function. Let $\Delta u \equiv u_H - u_L$ denote the payoff gain by observing a high signal instead of a low signal.

Proposition 4 (Forgetting Bad News) *Suppose that ψ_s is strictly convex, $s \in \{H, L\}$. Then in any PBE, $m_H^* > 0 \geq m_L^*$. Furthermore,*

$$\Delta u < \psi'_H(1 - \eta_H) \implies 0 < m_H^* < 1 - \eta_H \text{ and } m_L^* < 0.$$

If the marginal cost of remembering good news is lower than its marginal benefit for all $m_H \in [-\eta_H, 1 - \eta_H)$, i.e. $\psi'_H(1 - \eta_H) \leq \Delta u$, then the DM always remembers high signals. In this case, there is no point in trying to forget a low signal since the agent perfectly infers that a low signal was observed when she recollects $\hat{\sigma} = \varnothing$.

If the marginal cost of remembering good news is higher than its marginal benefit for some $m_H \in [-\eta_H, 1 - \eta_H)$, then the DM forgets high signals with positive probability. In this case, because the cost of a small amount of memory manipulation is of second-order, bad news are remembered with probability below the natural rate η_L, i.e., $m_L^* < 0$.

Next, I characterize the PBE in the forgetfulness model of Benabou and Tirole (Example 1) and in the limited memory model (Example 2) when signals have purely hedonic value.

The forgetfulness model of Benabou and Tirole (2002)

Consider the forgetfulness model of Example 1 and suppose that ψ_L is strictly convex. Given a low signal, self 1 solves

$$\max_{m_L \in [-1,0]} (1 + m_L) u_L - m_L \{\alpha (m_L^*, 0) u_H + [1 - \alpha (m_L^*, 0)] u_L\} - \psi_L (m_L) . \tag{1.7}$$

Applying Kuhn-Tucker's theorem and substituting the equilibrium condition $m_L = m_L^*$, we obtain

$$\frac{q\Delta u}{q - (1-q) m_L^*} = -\psi_L' (m_L^*) , \tag{1.8}$$

in any interior equilibrium.

Let m_L^* be implicitly defined by equation (1.8). From the implicit function theorem, such $m_L^* \in \mathbb{R}$ exists and is unique. The following proposition characterizes the PBE and presents some comparative statics results:

Proposition 5 (Characterization) *In the forgetfulness model when signals have a purely hedonic value, there exists an essentially unique PBE.*[25] *The equilibrium manipulation effort is*

$$m_L^* = \begin{cases} \psi_L'^{-1} \left(-\frac{q\Delta u}{q-(1-q)m_L^*} \right) & \text{if } \Delta u < -\frac{\psi_L'(-1)}{q}, \text{ and} \\ -1 & \text{if } \Delta u \geq -\frac{\psi_L'(-1)}{q}. \end{cases}$$

Furthermore, the absolute value of belief manipulation $|m_L^*|$ is:

1. *increasing in the benefit of manipulation Δu (for u_L fixed),*
2. *decreasing in the marginal cost of manipulation, and*
3. *increasing in q, the probability of not observing a signal.*

The comparative statics above follows from simple cost-benefit comparisons. When the marginal benefit of self-deception is higher or the marginal cost is lower, the agent chooses

[25] *The PBE is essentially unique in the sense that all PBE feature the same choices of actions a and b, the same manipulation efforts m_L and m_H, and the same beliefs for all recollections that are reached with positive probability. Equilibria may diverge only with respect to beliefs at recollections that are not reached with positive probability.*

to engage in more self-deception. This result is consistent with the experimental evidence presented by Prelec (2008), which suggests that self-deception is increasing in the benefits of manipulation.

Also, recall that in this model, no news is good news. Therefore, when the probability of not observing a signal q is higher, it becomes more credible that the individual has not manipulated her beliefs into forgetting a low signal. Hence, an increase in q increases the marginal benefit of self-deception, and this in turn leads to an increase in the amount of memory manipulation $|m_L^*|$.

The limited memory model

Consider the limited memory model of Example 2. Given a high signal, self 1 solves

$$\max_{m_H \in [-1,1]} m_H u_H + (1 - m_H) \left\{ \alpha \left(0, m_H^* \right) u_H + \left[1 - \alpha \left(0, m_H^* \right) \right] u_L \right\} - \psi_H \left(m_H \right).$$

Proceeding as in Proposition 5, it follows that the set of PBE efforts are characterized by

$$\frac{(1-q)\,\Delta u}{1 - q + q\left(1 - m_H^*\right)} = \psi_H' \left(m_H^* \right), \tag{1.9}$$

if $\Delta u \leq \psi_H' (1) \left[1 + \frac{q}{1-q} \left(1 - m_H^* \right) \right]$, and

$$m_H^* = 1 \text{ if } \Delta u \geq \psi_H' (1).$$

Since both sides of equation (1.9) are increasing in m_H^*, there may be multiple interior equilibria. It may also simultaneously feature interior equilibria and a corner equilibrium.[26] A person that believes she often forgets good signals is not hurt much by not recalling a good signal. Therefore, she will not manipulate her memory enough and, in equilibrium, she will often forget good signals. On the other hand, a person that usually remembers good signals is severely hurt by recollecting $\hat{\sigma} = \varnothing$. Therefore, she will have more incentive to remember good signals. As I show in the next section, these equilibria are welfare ranked (from an ex-ante

[26] For example, if $\psi_H' (1) \leq \Delta u \leq \psi_r' (1) \left[1 + \frac{q}{1-q} \left(1 - m_H^* \right) \right]$, there exist both an equilibrium with $m_H^* = 1$ and an interior equilibrium with m_H^* implicitly defined by equation (1.9).

perspective): The equilibrium with the lowest amount of memory manipulation is preferred. The individual may be caught in a self-trap where she exerts more manipulation effort because self 1 believes that she will have engaged in more memory manipulation.[27]

1.3 Purely Informative Signals and Information Acquisition

Suppose the decision-maker can choose whether or not to observe an informative signal. When would she prefer to observe it? This section is concerned with the implications of memory manipulation for the acquisition of information. I show that the DM will only observe a signal if the benefit of making an informed decision exceeds the cost of memory manipulation. Subsection 1.3.1 discusses a theory of regret aversion based on self-deception. Then, Subsection 1.3.2 shows that the model is consistent with intuitive behavior that Eliaz and Spiegler (2006) have shown to be incompatible with Caplin and Leahy's (2001) Psychological Expected Utility model.

The standard theory of information acquisition under expected utility states that it is optimal to observe a signal when the value of information (defined as the expected payoff gain by observing the signal) is greater than the cost of information. Similarly, I will show that the DM prefers to observe a signal if the (objective) value of information is greater than the expected cost of self-deception. In particular, when information has purely hedonic value, the DM always prefers not to observe any signal.

The *objective value of information* is defined as the expected payoff from observing the signal:

$$V \equiv E_{\hat{\sigma}_a}\left[E_\mu\left[u\left(a, b_a\left(\hat{\sigma}_a\right), \theta\right)|\hat{\sigma}\right]\right] - \max_{a\in A, b\in B}\int u\left(a, b, \theta\right) dF\left(\theta\right) \geq 0, \tag{1.10}$$

where $\max_{a\in A,b\in B}\int u\left(a,b,\theta\right)dF\left(\theta\right)$ is the expected payoff if the DM could not observe $\hat{\sigma}_a$. Thus, equation (1.6) implies that the ex-ante expected utility from observing the signal $U\left(\Sigma\right)$ is equal to

$$\max_{a\in A, b\in B}\int u\left(a, b, \theta\right) dF\left(\theta\right) + V - q\psi_H\left(m_H^*\left(a^*\right)\right) - \left(1-q\right)\psi_L\left(m_L^*\left(a^*\right)\right). \tag{1.11}$$

[27] The existence of multiple equilibria is interesting since there seems to be a large heterogeneity in the amount of self-deception accross different people (c.f., Prelec, 2008). However, since the results presented here hold in all PBE, they would also be obtained if one applied a selection criterion.

It follows that the DM would prefer to observe the signal if the objective value of information V is greater than the expected cost of memory manipulation $q\psi_H(m_H^*(a^*)) + (1-q)\psi_L(m_L^*(a^*))$.

Proposition 6 (Information Acquisition) *Fix a PBE. Let $U(\Sigma)$ denote the expected utility of observing the signal in this PBE and let $E[u]$ denote the expected utility of not observing the signal. Then, $U(\Sigma) - E[u] = V - q\psi_H(m_H^*(a^*)) - (1-q)\psi_L(m_L^*(a^*)) < V$.*

When information has a purely hedonic value, the objective value of information is $V = 0$. In equilibrium, when the a signal is forgotten ($\hat{\sigma} = \varnothing$), the DM knows that there is a probability $\alpha(m_L^*, m_H^*)$ that there was a high signal and $1 - \alpha(m_L^*, m_H^*)$ that there was a low signal. Bayesian updating implies that on average, the only effects of engaging in self-deception are the manipulation costs $\psi_L(m_L^*)$ and $\psi_H(m_H^*)$. Of course, there is still an interim incentive to manipulate beliefs after she observes the signal. *The inability to commit not to engage in self-deception leads to a loss in (ex-ante) expected utility*:

Corollary 1 *When information has purely hedonic value, the DM is strictly better off by not observing the signal: $E[u] > U(\Sigma)$. Furthermore, in order to observe the signal, the individual requires a "participation premium" of $q\psi_H(m_H^*) + (1-q) \times \psi_L(m_L^*)$.*

Proposition 6 and Corollary 1 show that memory manipulation leads to the avoidance of information when individuals have preferences over their own attributes (i.e., they have "ego utility").

The most standard model of ego utility one could formulate consists of a basic application of expected utility theory. Let the space of possible attributes Θ be a non-empty subset of $\mathbb{R}$ and let $F(.)$ denote the agent's prior distribution of θ. The DM has preferences that are represented by a strictly increasing von Neumann-Morgenstern utility function $u : \Theta \to \mathbb{R}$.

In this basic model, if the individual does not observe a signal that is informative about θ, her utility is $\int u(\theta)\, dF(\theta)$. If she observes a signal σ, the utility conditional on σ is $\int u(\theta)\, dF(\theta|\sigma)$. Hence, the expected utility of observing the signal is $\int_\sigma \int_\theta u(\theta)\, dF(\theta|\sigma)\, dG(\sigma)$, where G is the distribution of signals σ. By the law of iterated expectations, we have

$$\int u(\theta)\, dF(\theta) = \int_\sigma \int_\theta u(\theta)\, dF(\theta|\sigma)\, dG(\sigma),$$

so that *an individual with perfect memory and who behaves as an expected utility maximizer is always indifferent between observing the signal or not* when signals do not affect actions. In other words, in this standard model of ego utility, the fact that an individual has preferences over her expected attributes does not influence her decision of whether to acquire information. In particular, as in Blackwell's theorem, more information cannot hurt the individual.

Note that the result above holds regardless of the shape of the utility function u. In order to affect the decision of whether to acquire information, the utility function must be a non-linear function of *probabilities*. Several models of information acquisition have, thus, assumed that utility functions are non-linear in probabilities.[28] Our model also leads to a utility function that is non-linear in probabilities. However, the non-linearity arises endogenously through memory manipulation. Therefore, the present model can be seen as providing a cognitive foundation for a model of information acquisition.

Proposition 6 shows that the DM will prefer not to collect some information if its objective value V is lower than the expected costs from memory manipulation $q\psi_H\left(m_H^*\left(a^*\right)\right) + \left(1-q\right)\psi_L\left(m_L^*\left(a^*\right)\right)$. In particular, she will always prefer not to observe information that is informative about her attributes θ but does not affect her actions b. For example, people will prefer not to know the outcome of a medical exam if the value of information is not sufficiently high (e.g. if a detected disease is not treatable) and if the exam has a potentially large impact on the person's self-image. Dawson et al. (2006) present experimental evidence supporting this result.[29]

An immediate consequence of avoiding information correlated with one's skills is the possible desirability of "self-handicapping" strategies such as under-preparing for an examination or getting too little sleep before a physical exercise (Berglas and Baumeister, 1993). Self-handicapping strategies reduce the informational content of the signal, and therefore, the model predicts that a person may engage in such strategies if the expected costs are not too high.

In several environments, competition allows for more precise information about one's abil-

[28] For example, Philipson and Posner (1995) and Caplin and Eliaz (2003) analyze the case of testing for sexually transmitted diseases, Kőszegi (2003) considers a model of patient decision-making, Kőszegi (2006) studies information acquisition and financial decisions, and Caplin and Leahy (2004) study strategic information transmission. With the exception of Philipson and Posner (1995), who do not provide a justification for the assumption of a utility function that is non-linear in probabilities, all these papers depart from the standard expected utility model by adopting the Psychological Expected Utility model.

[29] Dunning (2005) obtained the same result in the domain of academic ability.

ities. Thus, individuals may display a "fear of competition" and prefer environments where outcomes are not directly comparable to the outcomes from other people. More generally, the model predicts that in environments where information is correlated with one's attributes, individuals typically face a trade-off between the objective value of information and the costs of self-deception. Coarser information structures reduce the objective value of information but cause lower self-deception costs.

1.3.1 Regret Aversion

In this subsection, I study how the agent's utility from the lottery changes as a function of her prior distribution about her attributes. This allows us to show that the model developed in this essay provides a formalization for the (informal) theory of regret aversion based on self-evaluation proposed by Josephs et al. (1992).

The Theory of Regret Aversion based on Self-Perceptions The theory of choice based on regret aversion was simultaneously proposed by Bell (1982) and Loomes and Sugden (1982). According to this theory, agents base their decisions not only on expected payoffs but also on the payoffs that they would have obtained if they had made other decisions. Because agents anticipate feeling regret or delight over their choice, they take this into account when making a decision.

Josephs et al. (1992) argued that the feeling of regret arises from an individual's self-evaluation that follows an outcome.[30] They suggested that people with worse self-perceptions are more severely harmed by negative outcomes than those with better self-perceptions. Therefore, *individuals with low self-image would be more likely to make choices that minimize the possibility of regret.*

According to this theory of regret aversion based on self-perception, the premium required to observe a signal σ_a that is informative about the DM's attributes should be decreasing in the favorableness of the agent's prior distribution (see Figure 1-8). Denote by $U(\Sigma_a)$ the expected utility of observing signal σ_a and, as in Proposition 6, let $E[u]$ denote the expected utility from not observing the signal. Then, the theory predicts that $E[u] - U(\Sigma_a)$ should be decreasing in

[30] See also Larrick (1993) for a similar discussion.

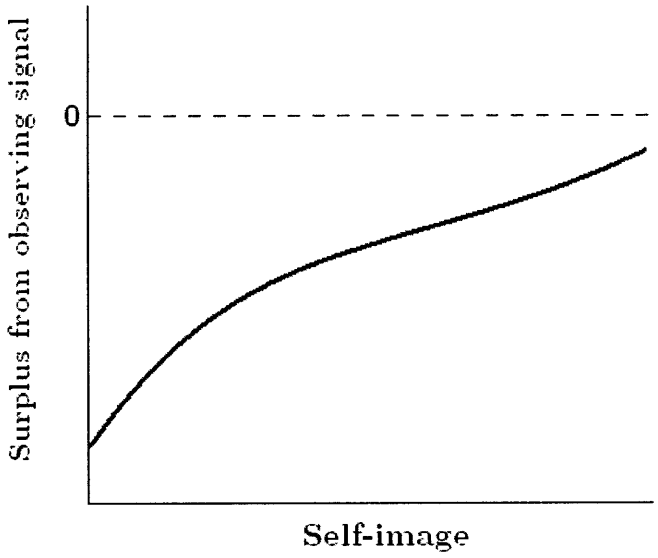

Figure 1-8: Regret Aversion based on Self-Perceptions

the favorableness of the agent's prior distribution over her attributes.

The Model Since the theory presented by Josephs et al. (1992) considers only choices where no ex-post actions are taken, assume that B is a singleton. Moreover, since the ex-ante decision consists of selecting a gamble, we interpret ex-ante actions $a \in A$ as a choice between different possible lotteries and assume that these actions do not affect the DM's utility function. The only way in which ex-ante actions $a \in A$ affect the agent's utility is through the different distributions associated with each lottery. For simplicity, I consider either the forgetfulness model of Example 1 or the limited memory model of Example 2.

In order to determine how the agent's attitude toward information is affected by her prior, let κ be a parameter that indexes her prior distribution. A higher parameter κ leads to a more favorable prior in the sense of first-order stochastic dominance:

$$\kappa' > \kappa \implies F\left(\theta;\kappa'\right) \leq F\left(\theta;\kappa\right), \tag{1.12}$$

for all $\theta \in \Theta$, with strict inequality for some θ.

Denote the gain from observing a high signal instead of a low signal by

$$\Delta u(\kappa, a) = \int u(\theta)\, dF(\theta|\sigma_a = H; \kappa) - \int u(\theta)\, dF(\theta|\sigma_a = L; \kappa).$$

The assumption that individuals with worse self-perceptions are more severely harmed by negative outcomes than those with better self-perceptions states can be stated as:[31]

Assumption 3. $\Delta u(\kappa, a)$ is decreasing in κ for all $a \in A$.

Recall that $U(\Sigma_a)$ and $E[u]$ were defined as the expected utility of observing signal and the expected utility from not observing the signal, respectively. Then, the prediction of the theory of regret aversion based on self-perceptions can be stated as follows:

Conjecture 1 (Josephs et al., 1992) *$E[u] - U(\Sigma_a)$ is positive and decreasing in κ, for all $a \in A$.*

Next, I show that under Assumption 3, the model implies that Conjecture 1 is true. Since there are no ex-post actions in this model, the only benefit from memory manipulation is the change in the DM's self-perceptions Δu. Therefore, the amount of memory manipulation is increasing in the self-image gain from observing a high signal Δu. Because, under Assumption 3, $\Delta u(\kappa, a)$ is decreasing in κ, we obtain:

Proposition 7 (Regret Aversion) *Suppose Assumption 3 holds and consider either the forgetfulness model of Example 1 or the limited memory model of Example 2. For any $a \in A$, the premium required to observe the signal σ_a is decreasing (in the sense of strong set order) in the the decision-maker's prior over her attributes indexed by parameter κ.*

Therefore, the model provides a formalization of the theory of regret aversion based on self-perception proposed by Josephs et al. (1992).

[31] Although this assumption is intuitive for intermediate values of κ, it is probably not a good assumption for extreme values of κ. An individual who is certain to have a very high parameter θ is likely to be severely hurt by a negative outcome. Similarly, a positive outcome may have a large effect on the self-perceptions of someone who previously believed to have an extremely low θ.

1.3.2 Prior-Dependent Attitude Towards Information

Proposition 6 showed that the DM will seek information if its objective value is greater than the expected costs of self-deception. This result contrasts with Blackwell's theorem, which states that more information cannot be harmful. Alternatively, Caplin and Leahy (2001) have proposed the Psychological Expected Utility (PEU) model which generalizes the expected utility model to allow for different attitudes towards information.

Eliaz and Spiegler (2006) have criticized the PEU model by showing that it is inconsistent with certain situations where a DM's preference for information varies with her prior distribution. In one example, they describe a patient who prefers more accurate medical tests when she is relatively certain of being healthy, yet she avoids these tests when she is relatively certain of being ill. In another example, they describe a manager that asks for their employees' opinion only when he is sufficiently certain that the new information will not cause her to change her views much. They proved that such behaviors are inconsistent with the PEU model. As a result, Eliaz and Spiegler have suggested that one should drop the Bayesian updating assumption.

As the following example shows, the model presented in this essay is consistent with these two examples described by Eliaz and Spiegler. Therefore, unlike the PEU model, the self-deception model leads to prior-dependent attitudes toward information while retaining Bayesian updating.[32]

Example 3 *An individual must choose whether or not to take some medical exam. Let* $a = E$ *denote the choice of taking the exam and* $a = NE$ *denote the choice of not taking it. The exam is informative about the individual's health* θ *and has outcome* $\sigma = H$ *if the individual is healthy and* $\sigma = L$ *if she is not. If the individual takes the exam, she can undertake medical treatment* $B = \{T, NT\}$, *where* $b = T$ *and* $b = NT$ *denote the cases where she does and does not undertake the treatment.*

The individual's payoff from being healthy is 25. *If she takes the medical exam, the individual has a cost of* 5. *Thus, her payoff conditional on a high signal is* 25 *if* $b = T$ *and* 20 *if* $b = NT$. *The agent's expected payoff conditional on a low signal is* $\gamma(q)$. *Undertaking the treatment can*

[32] Epstein (2007) presents a model of anticipatory utility. In the special case of rank-dependent expected utility, they show that their model is also able to accommodate the behavior from Eliaz and Spiegler's examples.

reduce the effects from the disease, which increases her expected payoff to $\gamma(q)+1$. *In order to be consistent with Assumption 3, assume that* $\gamma(q)$ *is increasing so that* Δu *is decreasing in the DM's prior distribution over her skills (indexed by the probability of observing a high signal* q*). Let* $\gamma\left(\frac{1}{2}\right)=10$ *and* $\gamma(1)=20$. *If the DM does not take the exam, she obtains an expected payoff of* $25q+\gamma(q)(1-q)$.

For simplicity, let the memory system be given by the forgetfulness model of Example 1 and suppose that memory manipulation is binary: $m_L \in \left\{-\frac{1}{2},0\right\}$ *with* $\psi_L\left(-\frac{1}{2}\right)=3$. *The decision problem is depicted in Figure 1-9.*

It is straightforward to show that the agent chooses $m_L=0$ *when* q *is close to* 1. *In this case, since the objective value of information is positive and the cost of self-deception is zero, the DM always chooses to take the exam. When* $q=\frac{1}{2}$, *however, the DM engages in self-deception (*$m_L=-\frac{1}{2}$*). It can be shown that the expected cost from memory manipulation is greater than the objective value of information so that the DM prefers not to take the exam. Thus, unlike Eliaz and Spiegler's result on the PEU model, the DM may prefer to take the exam when she is relatively certain of being healthy (*$q\approx 1$*) but prefer not to take the exam for intermediate values of* q.

1.4 Lotteries Over Money

> We propose that the consequences of each bet include, besides monetary payoffs, the credit or blame associated with the outcome. Psychic payoffs of satisfaction or embarrassment can result from self-evaluation or from an evaluation by others. (Heath and Tversky, 1991, pp. 7-8)

In Section 1.3, outcomes $\sigma \in \{H, L\}$ consisted of purely informative signals, which affected the DM's utility only through her beliefs about her own attributes θ. This section considers outcomes that affect the DM's utility not only by providing information about θ but also directly through monetary payments. I show that the model leads to a theory of ambiguity aversion based on self-deception. The DM may reject gambles with small but positive expected value. Moreover, the model is consistent with the competence hypothesis proposed by Heath and Tversky (1991).

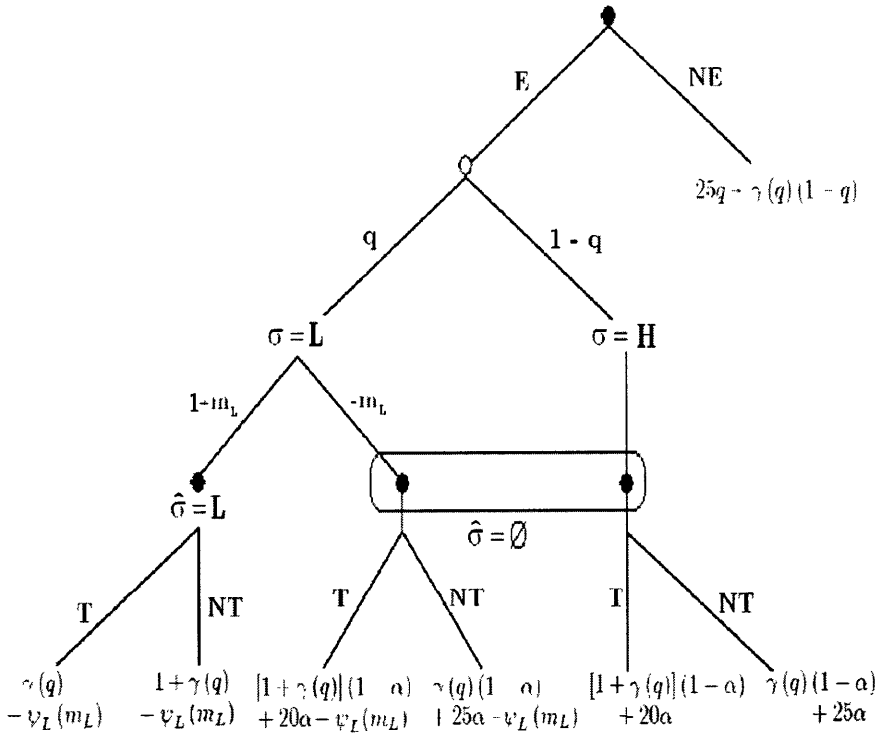

Figure 1-9: Decision Problem in Example 3

In order to focus on the implications of the model for the DM's preferences over monetary lotteries, I take A and B to be singletons so that the agent does not take any actions. Therefore, as in the model of Subsection 1.2.3, information has purely hedonic value. However, in the case of monetary lotteries, outcomes also have a direct effect on the DM's payoff through monetary payments.

As described in Section 1.2, the outcome $\sigma \in \{L, H\}$ is interpreted as a monetary payment. For notational simplicity, I omit a and b from the DM's von Neumann-Morgenstern utility function. Therefore, in this section, the DM's utility function is denoted by $u(\theta, x)$, where $x \in \mathbb{R}$ denotes the amount of money that she has. If $H > L$, a high outcome not only provides favorable information about the agent's attributes θ but also leads to a higher monetary payment. This is the natural assumption since, in most cases, the outcome associated with higher monetary payments is also associated with better attributes. If $L > H$, a high outcome provides favorable information about θ but provides a lower payment. The results in this essay hold for any L

and H.

For simplicity, I assume that the utility function is additively separable between characteristics and money:

$$u(\theta, x) = v(\theta) + \tau(x),$$

for a strictly increasing function $v : \Theta \to \mathbb{R}$ and a function $\tau : \mathbb{R} \to \mathbb{R}$. Appendix A analyzes the general case. Let v_s denote the expected payoff from attributes conditional on recollection $\hat{\sigma} \in \{H, L, \varnothing\}$.

Under additive separability, monetary payments can be factored out of self 1's memory manipulation choice. Given an outcome $\sigma = s \in \{H, L\}$, she maximizes:

$$(\eta_s + m_s) v_s + (1 - \eta_s - m_s) v_\varnothing + \tau(s) - \psi_s(m_s).$$

Therefore, self 1 chooses the same amount of memory manipulation as in the purely hedonic signals model analyzed in Subsection 1.2.3. Proposition 4 then implies that the DM will never choose to remember a low outcome or forget a high outcome:

Corollary 2 *Suppose that ψ_s is strictly convex, $s \in \{H, L\}$. Then, in any PBE, $m_H^* > 0 \geq m_L^*$. Furthermore,*

$$v_H - v_L < \psi_H'(1 - \eta_H) \implies 0 < m_H^* < 1 - \eta_H,\ m_L^* < 0.$$

From equation (1.6), the DM's ex-ante expected utility is:

$$U(\Sigma) = q[v_H + \tau(H) - \psi_H(m_H^*)] + (1 - q)[v_L + \tau(L) - \psi_L(m_L^*)]. \tag{1.13}$$

It consists of the sum of the expected payoff from attributes, the expected monetary payoffs, and the expected cost of memory manipulation. Denote by U^I the utility of a lottery with the same distribution over monetary outcomes as the one above but whose monetary outcomes are uninformative about θ. Then, the DM's ex-ante expected utility can be written as

$$U(\Sigma) = U^I - q\psi_H(m_H^*) - (1 - q)\psi_L(m_L^*). \tag{1.14}$$

Because the DM takes no actions after observing the outcome (i.e., B is a singleton), information

has no objective value. Therefore, the model implies that the uninformative lottery is strictly preferred.

Remark 2 *Consider the entrepreneurship model described in Subsection 1.2. The DM will choose to become an entrepreneur if the expected monetary payoffs are greater than the expected costs of self-deception:*

$$q\tau(H) + (1-q)\tau(L) \geq q\psi_H(m_H^*) + (1-q)\psi_L(m_L^*).$$

Baron (1999) presents evidence that individuals who become entrepreneurs find it easier to admit past mistakes to themselves. In a static environment, our model may easily lead to this result. Suppose, for example, that individuals have heterogeneous concerns for self-image or that homogeneous individuals play different equilibria of the game. Then those with a lower concern for self image or those who play equilibria with lower amounts of self-deception are precisely the ones who benefit the most from becoming entrepreneurs. Alternatively, Section 1.5 will establish that the expected cost of self-deception converges to zero as experience grows. Therefore, it could be the case that entrepreneurs were not different from other individuals ex-ante, but, as they have gained experience, their cost of admitting past mistakes decreased.

Remark 3 *Consider the used car model described in Subsection 1.2. The individual will purchase the car if the expected payoff gain from the purchase is greater than the expected costs of forgetting a bad outcome.*

Remark 4 *Under the additive separability assumption, it is immediate to extend Proposition 7 to the case of monetary lotteries. Let κ index the DM's prior distribution as defined in equation (1.12). As in Assumption 3, assume that $\Delta v(\kappa)$ is decreasing in κ and consider either the forgetfulness model of Example 1 or the limited memory model of Example 2. Then, the premium $U^I - U(\Sigma)$ is positive and decreasing (in the sense of strong set order) in κ.*

1.4.1 Probability Weights

In this subsection, I will consider a non-expected utility representation, where the decision-maker's expected utility from observing the signal is expressed as a weighted average of the

utility in each state of the world $\sigma \in \{L, H\}$. The representation consists of a weighting function $w : [0,1] \to \mathbb{R}$ such that the utility from participating in the lottery is

$$U(\Sigma) = w(q) \times u_H + [1 - w(q)] \times u_L,$$

where $u_s \equiv \int u(\theta, s)\, dF(\theta|\sigma = s)$, $s \in \{H, L\}$. Clearly, the decision maker is an expected utility maximizer if $w(q) = q$. Although the model does not feature ambiguity in the sense of an imprecise distribution of probabilities, I will follow the literature on decision-making under ambiguity and say that an agent is *ambiguity averse* if $w(q) < q$.[33]

Proposition 8 shows that the ex-ante preferences of the DM can be represented by a non-expected utility and that the DM always displays ambiguity aversion when the outcomes from the lottery are informative about her attributes:[34]

Proposition 8 (Representation) *The DM's expected utility from the monetary lottery can be represented by*

$$U(\Sigma) = w(q)\, u_H + [1 - w(q)]\, u_L, \tag{1.15}$$

where

$$w(q) = q - \frac{q\psi_H(m_H^*) + (1-q)\,\psi_L(m_L^*)}{u_H - u_L}. \tag{1.16}$$

Furthermore, $w(0) = 0$, $w(1) = 1$, *and* $w(q) < q$ *for all* $q \in (0,1)$.

Remark 5 *Note that the representation from equation (1.15) is not separable between probabilities and the utility* u_s. *Since the departure from linear probability weights is caused by memory manipulation, individuals who engage in more memory manipulation have lower probability weights* $w(q)$. *Furthermore, because the amount of memory manipulation is increasing in the marginal utility from attributes, it follows that the deviation from linear weighting is itself a function of* u_s.

[33] If we identify "unambiguous" lotteries as those whose outcomes are uninformative about the DM's attributes and follow the approach in Epstein (1999), it follows that the DM is ambiguity averse if and only if $w(q) < q$.

[34] Note that the model of monetary lotteries becomes a model of purely hedonic signals (Subsection 1.2.3) when $\tau(H) = \tau(L)$. Thus, when information has purely hedonic value, the DM's expected utility from observing the signal can be represented by

$$U(\Sigma) = w(q)\, u_H + [1 - w(q)]\, u_L,$$

where $w(q) = q - \frac{q\psi_H(m_H^*) + (1-q)\psi_L(m_L^*)}{u_H - u_L}$. Furthermore: $w(0) = 0$, $w(1) = 1$, and $w(q) < q$ for all $q \in (0,1)$.

1.4.2 Discussion

The model presented here implies that ambiguity aversion is a consequence of the lottery outcomes being informative about the DM's attributes. Several experimental papers have related ambiguity aversion with the lotteries' being influenced by an individual's skill or knowledge.[35] First, some experiments have contradicted the idea that ambiguity aversion is related to the imprecision of the probability distribution of the events as is usually argued. Budescu, Weinberg, and Wallsten (1988), for example, compared decisions based on numerically, graphically (the shaded area in a circle), and verbally expressed probabilities. Numerical descriptions of a probability are less vague than graphic descriptions which, in turn, are less vague than verbal descriptions. Thus, if agents had a preference for more precise distributions, they should rank events whose probabilities have a numerical description first, graphic descriptions second, and verbal descriptions last. However, unlike ambiguity aversion would predict, subjects were indifferent between these lotteries. Indeed, the authors could not reject that the agents behaved according to subjective expected utility theory and weighted events linearly.[36]

Heath and Tversky argued that people's preferences over ambiguous events arise from the anticipation of feeling knowledgeable or competent.[37] Their interpretation of the Ellsberg paradox is as follows:

> People do not like to bet on the unknown box, we suggest, because there is information, namely the proportion of red and green balls in the box, that is knowable in principle but unknown to them. The presence of such data makes people feel less knowledgeable and less competent and reduces the attractiveness of the corresponding bet. (Heath and Tversky, 1991, pp. 8)

Fox and Tversky (1995, pp. 585) proposed that ambiguity is caused by comparative ignorance. They have argued that "ambiguity aversion is produced by a comparison with less ambiguous events or with more knowledgeable individuals." As in Heath and Tversky's (1991) competence hypothesis, this "comparative ignorance hypothesis" states that ambiguity aver-

[35] See Goodie and Young (2007) for a detailed discussion of this literature.

[36] See also Budescu et al. (2002).

[37] Subsection 1.4.4 defines Heath and Tversky's "competence hypothesis" more precisely and also briefly reviews the empirical evidence related to it.

sion is driven by the feeling of incompetence. Similarly, Goodie (2003) proposed the *perceived control* hypothesis, according to which ambiguity aversion is generated by an agent's belief that the distribution of outcomes is influenced by attributes such as knowledge or skill.[38]

As will be shown in Section 1.5, it is straightforward to embed the model in a dynamic setting where the DM updates beliefs according to Bayes' rule. Therefore, the model provides a tractable framework where individuals display ambiguity aversion and still follow Bayes' rule. Under the self-perception reinterpretation of ambiguity aversion, the difficulties in characterizing an updating rule under ambiguity do not arise.[39]

1.4.3 Small-Stakes Risk Aversion

This subsection considers lotteries with small monetary stakes. It is shown that memory manipulation leads the DM to exhibit "zeroth-order" risk aversion, which has important implications. Standard expected utility maximizers exhibit second-order risk aversion. An individual with second-order risk aversion always accepts small gambles with positive expected value. Then, if the agent has reasonable levels of risk aversion with respect to lotteries with small stakes, she must display unrealistically high levels of risk aversion with respect to lotteries with large stakes (Samuelson, 1963; Rabin, 2000).

Segal and Spivak (1990) show that an individual with first-order risk aversion rejects small gambles as long as the positive expected value is sufficiently small. Therefore, several nonexpected utility models that feature first-order risk aversion have been proposed. However, Safra and Segal (2008) show that the inability to simultaneously explain an agent's risk aversion over lotteries with small stakes and lotteries with large stakes can be generalized to non-expected utility models.[40] In this subsection, I show that the model allows us to reconcile risk aversion with respect to small lotteries with sensible levels of risk aversion with respect to large lotteries.[41]

[38] There is a large experimental literature on the effect of perceived control on risk-taking (c.f., Chau and Phillips, 1995, or Horswill and McKenna, 1999).

[39] See, for example, Hanany and Kilbanoff (2007).

[40] The crucial assumption in Segal and Spivak (1990) is that decision-makers have a unique preference relation over final-wealth distributions. Hence, their framework does not include gain-losses models such as Prospect Theory (Kahneman and Tversky, 1979; Tversky and Kahneman, 1992).

[41] Fudenberg and Levine (2007) present a dual-self model of dynamic consumption where the agent's exercise of self-control may also lead to risk aversion over lotteries with small stakes.

Consider the lottery described previously. The certainty equivalent of the lottery is defined by the monetary amount $CE \in \mathbb{R}$ that makes the agent indifferent between participating in the lottery or receiving CE for sure:

$$\int u(\theta, CE)\, dF(\theta) = qu_H(H) + (1-q)\, u_L(L) - q\psi_H(m_H^*) - (1-q)\,\psi_H(m_L^*). \tag{1.17}$$

The risk premium associated with a lottery is defined as the difference between the expected payment and the certainty equivalent: $\pi = qH + (1-q)\, L - CE$.

Let $s \in \{H, L\}$ be a binary random variable such that $E[s] = qH + (1-q)\, L = 0$. Consider a lottery that pays $x = \varepsilon s$, where $\varepsilon > 0$. A decision maker has risk preferences of second order if $\lim_{\varepsilon \to 0_+} \pi(\varepsilon)/\varepsilon = 0$. She is first-order risk averse if $\lim_{\varepsilon \to 0_+} \pi(\varepsilon)/\varepsilon > 0$ is finite. She is zeroth-order risk averse if $\lim_{\varepsilon \to 0_+} \pi(\varepsilon)/\varepsilon = +\infty$.

Note that the monetary lottery converges to a model of purely hedonic signals as ε approaches zero. Then, as shown in Corollary 1, the DM demands a strictly positive participation premium in order to observe the signal. Hence, the certainty equivalent of the lottery converges to $CE(0) < 0$ and

$$\lim_{\varepsilon \to 0_+} \frac{\pi(\varepsilon)}{\varepsilon} = -\lim_{\varepsilon \to 0_+} \frac{CE(\varepsilon)}{\varepsilon} = +\infty.$$

Thus, the individual exhibits zeroth-order risk aversion. This result is established formally in the following proposition:

Proposition 9 (Zeroth-Order Risk Aversion) *In any PBE, the DM exhibits zeroth-order risk aversion.*

Since outcomes are informative about the DM's attributes, the DM engages in memory manipulation. Therefore, even when the monetary payoffs converge to zero, she still demands a strictly positive risk premium. Hence, the individual displays zeroth-order risk aversion and displays risk aversion for lotteries with small stakes. However, as shown in Corollary 1, when the expected monetary stakes are larger than the DM's participation premium, she will accept to participate in the lottery. Thus, as the following example shows, the DM may be risk averse over lotteries with small stakes without displaying an unreasonable degree of risk aversion over lotteries with large stakes:

Example 4 (Safra and Segal, 2008) *Suppose an agent rejects a lottery that pays either* -100 *or* 105 *with equal probability at all wealth levels below* $300,000$. *Safra and Segal show that all standard non-expected utility models imply that this agent cannot accept a lottery that pays* $-5,000$ *or* $10,000,000$ *with equal probability for some wealth level below* $300,000$. *The model is this essay, however, is consistent with this behavior. Indeed, I will show that the DM may even accept the second lottery for all wealth levels below* $300,000$.

Suppose both lotteries have the same informational content about the DM's attributes θ. *For simplicity, take the forgetfulness model of Example 1 with binary manipulation efforts* $m_L \in \left\{-\frac{1}{2}, 0\right\}$ *and let* $\tau(x) = x$ *for all* $x \in \mathbb{R}$. *Suppose that* $\frac{1}{3}(v_H - v_L) > \psi_L\left(-\frac{1}{2}\right)$ *so that self 1 engages in memory manipulation:* $m_L^* = -m$. *Then, the DM rejects the first lottery and accepts the second lottery for all wealth levels below* $300,000$ *if*

$$\frac{1}{2}(v_L + W - 100) + \frac{1}{2}(v_H + W + 105) - \frac{1}{2}\psi_L\left(-\frac{1}{2}\right) < \frac{1}{2}v_L + \frac{1}{2}v_H + W, \text{ and}$$

$$\frac{1}{2}(v_L + W - 5000) + \frac{1}{2}(v_H + W + 1000000) - \frac{1}{2}\psi_L\left(-\frac{1}{2}\right) > \frac{1}{2}v_L + \frac{1}{2}v_H + W$$

for all $W \leq 300,000$. *These conditions are satisfied if*

$$5 < \psi_L\left(-\frac{1}{2}\right) < \min\left\{\frac{1}{3}(v_H - v_L); 995,000\right\}. \tag{1.18}$$

Therefore, when inequality (1.18) is satisfied, the DM accepts the first lottery and rejects the second lottery for all wealth levels below $300,000$.

1.4.4 The Competence Hypothesis

Consider two lotteries with the same distribution over monetary outcomes. In the first lottery, outcomes are informative about the decision-maker's skills or knowledge whereas in the second they are not. If the information about one's skills or knowledge is not useful (i.e., the objective value of information from the first lottery is zero) and the individual is an expected utility maximizer, she should be indifferent between these lotteries. Since one's attributes are ambiguous, an ambiguity averse individual should prefer the lottery whose outcomes are uninformative about her skills or knowledge.

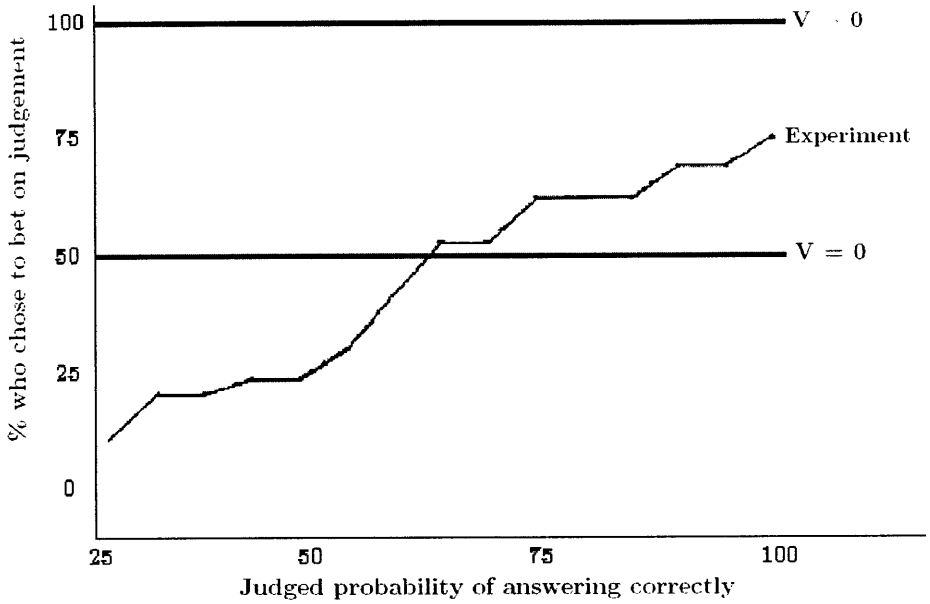

Figure 1-10: Experimental Evidence on the Competence Hypothesis

Heath and Tversky (1991) have studied this choice in a series of experiments. They have shown that people prefer the skill- or knowledge-dependent lottery in contexts where they feel knowledgeable or competent but prefer the skill- or knowledge-independent lottery in ones where they consider themselves ignorant or uninformed. In one experiment, for example, subjects were asked to answer several questions. Subjects also revealed (in an incentive-compatible way) their expected probability of answering the questions correctly. Afterwards, they chose between betting on their answers or participating in a lottery with the same expected probability of winning. The proportion of people who chose to bet on their answers is presented in Figure 1-10.

If decision-makers are expected utility maximizers and the value of information is zero, they should be indifferent between these two lotteries. Since Prospect Theory does not distinguish between sources of uncertainty in the specification of probability weighting function, it also predicts that people should be indifferent between these two lotteries. Therefore, in both cases, the proportion of individuals who bet on the knowledge-based lottery should be roughly constant at 50% when the value of information is zero. If the value of information is positive,

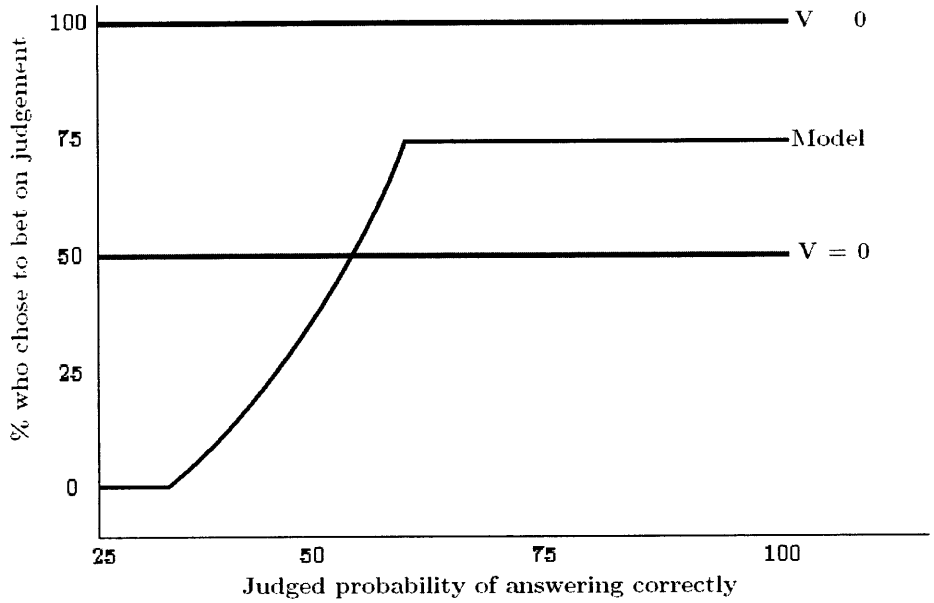

Figure 1-11: Simulated Results on the Competence Hypothesis

individuals who behave according to either Expected Utility Theory or Prospect theory should prefer the knowledge-based lottery. Hence, in this case, the proportion of individuals who bet on the knowledge-based lottery should be constant at 100%.

Heath and Tversky found a remarkably different pattern. The proportion of people who preferred to bet on the knowledge-based lottery instead of a knowledge-independent lottery *with the same expected probability of winning* was increasing in the judged probability.[42] In situations where the expected probability of winning was small, people preferred to bet on the knowledge-independent lottery. On the other hand, when the expected probability of winning was large, individuals preferred to bet on the knowledge-based lottery. This result was labeled the Competence Hypothesis.

The following examples show that our model is consistent with the Competence Hypothesis:

Example 5 *Consider the forgetfulness model of Example 1. For simplicity, let memory manipulation be a binary variable* $m_L \in \left\{-\frac{2}{3}, 0\right\}$, *with* $\psi_L\left(-\frac{2}{3}\right) = \frac{1}{4}$ *and* $\psi_L(0) = 0$. *Suppose that*

[42] A number of other experiments have confirmed the predictions of the competence hypothesis (c.f., Keppe and Weber, 1995; Taylor, 1995; Kilka and Weber, 2000; Chow and Sarin, 2001; Fox and Weber, 2002; Kuehberger and Perner, 2003; Di Mauro, 2008).

the DM does not face any ex-ante choice. However, in order to have a positive objective value of information, suppose that she chooses between a low b_L and a high b_H action ex-post.

Let $v_s(b)$ denote the expected payoff from attributes conditional on outcome $s \in \{H, L\}$ and take the following payoffs:

$$v_H(b_H) = 6,\ v_H(b_L) = 5,\ v_L(b_L) = 1,\ v_L(b_H) = 0,\ \tau(H) = 1,\ \tau(L) = 0.$$

In Appendix D, I show that the DM prefers the attribute-dependent lottery if $q > \frac{11}{23}$ and prefers the attribute-independent lottery if $q < \frac{11}{23}$.

Example 6 *Take the same parameters from the previous example but suppose that memory manipulation is a binary variable $m_L \in \{-\bar{m}, 0\}$, where the parameter $\bar{m}$ is distributed according to a c.d.f. Φ on $\left[\frac{1}{3}, 1\right]$. Suppose that $q \geq \frac{1}{4}$ so that the objective value of information is positive.*[43] *In Appendix D, I show that the proportion of people who prefer the attribute-dependent lottery is:*

$$\begin{gathered} 0 \text{ if } q \in \left[\frac{1}{4}, \frac{7}{19}\right], \\ \Phi\left(\frac{3}{4} - \frac{1-2q}{1-q}\right) \text{ if } q \in \left(\frac{7}{19}, \frac{1}{2}\right], \\ \Phi\left(\frac{3}{4}\right) \text{ if } q \in \left(\frac{1}{2}, 1\right]. \end{gathered}$$

Therefore, consistent with the Competence Hypothesis, this proportion in increasing in q. Figure 1-11 depicts the case where $\bar{m}$ is uniformly distributed.

1.5 Practice makes perfect: The Repeated Model

The previous sections considered a decision-maker who observes an outcome once and makes inferences about her attributes based on her recollection of this outcome. In several situations, however, individuals participate in this process repeatedly. A professional investor, for example, is constantly deciding which investment to undertake and receives feedback about the success

[43] *If $q < \frac{1}{4}$, then $b(\hat{\sigma}) = b_L$ for all $\hat{\sigma}$. Therefore, since actions are not a function of recollections, the objective value of information is zero.*

or failure of these investments very frequently. It is often argued that the biases in decision-making that we observe in experimental settings would be severely attenuated as individuals gain experience. This section presents a repeated version of the general model described in Section 1.2 and shows that this is indeed the case in this model. More precisely, I show that the behavior of the DM converges to the one predicted by expected utility theory as the number of observed signals grows.

Consider a repeated version of the general model described in Section 1.2. For simplicity, I assume that A is a singleton so the DM only chooses actions after observing the signal.[44] In each period $n \in \{1, 2, 3, ..., N\}$, an independent draw of the signal $\sigma_n \in \{H, L\}$ is made. Each signal σ_n is observed with probabilities $\Pr(\sigma = H|\theta)$ and $\Pr(\sigma = L|\theta)$, where θ is the agent's 'true' attributes. The parameter θ is not known. Instead, the DM has a prior $F(\theta)$ about its distribution. Hence, the prior over the distribution of a signal σ_n is

$$\Pr(\sigma_n = s) = \int \Pr(\sigma_n = s|\theta)\, dF(\theta) \qquad s \in \{H, L\},$$

where the conditional probability $\Pr(\sigma = H|\theta)$ is strictly increasing in θ.

After observing $\sigma_n \in \{H, L\}$, the DM engages in memory manipulation m_L and m_H. She recollects a signal $\hat{\sigma}_n \in \{H, L, \varnothing\}$. A history at time n is a sequence of recollected signals and actions:

$$h^{n-1} = ((\hat{\sigma}_1, ..., \hat{\sigma}_{n-1}), (b_1, ..., b_{n-1})) \in H^{n-1},$$

where $H^{n-1} \equiv \{\varnothing, L, H\}^{n-1} \times B^{n-1}$ is the set of possible histories. Note that, in this model, the DM can only manipulate the recollection of a signal in the time that the signal occurred. After the recollection has been registered into the agent's memory, she can no longer distort it.[45]

As in the static game, the agent's choice is modeled through a different self acting each time information is forgotten. Thus, in each period, a stage-1 self chooses memory manipulations

[44] See Remark 6.

[45] This assumption captures the psychological finding that most information loss occurs soon after it is obtained. Nevertheless, it is clearly an extreme assumption. In general, forgetting rates seem to follow a power law (Anderson, 1995). Therefore, a large fraction of the information is lost right after learning, and over time, the rate of forgetting slows down.

$(m_{H,n}, m_{L,n}) : H^{n-1} \to [-\eta_H, 1-\eta_H] \times [-\eta_L, 1-\eta_L]$ to maximize the discounted sum of payoffs from all future stage-games. The discount rate is $\delta \in [0,1)$. Then, a stage-2 self applies Bayes' rule and chooses an action $b_n : \{\varnothing, L, H\} \times H^{n-1} \to B$. For notational clarity, I omit the arguments from the profiles of actions and manipulations.[46]

Definition 2 *A PBE of the game is a strategy profile* (b^*, m_H^*, m_L^*) *and posterior beliefs* $\mu(.|.)$ *such that:*

1. $m_{s,n}^*$ *maximizes*

$$\begin{aligned}&(\eta_s + m_s)\left\{E_\mu\left[u(\theta, b_n^*, s) \mid \left(s, b(s); h^{n-1}\right)\right] + \delta V\left(s, b(s); h^{n-1}\right)\right\}\\&+ (1-\eta_s - m_s)\left\{E_\mu\left[u(\theta, b_n^*, s) \mid \left(\varnothing, b(\varnothing); h^{n-1}\right)\right] + \delta V\left(\varnothing, b(\varnothing); h^{n-1}\right)\right\} - \psi_s(m_s)\end{aligned}$$

   with respect to m_s, $s \in \{H, L\}$.

2. $b_n^* \in \arg\max_{b \in B}\left\{E_\mu\left[u(\theta, b, s) \mid \hat{s}, h^{n-1}\right] + \delta V\left(\hat{s}, b; h^{n-1}\right)\right\}$, *for* $s \in \{H, L\}$ *and* $\hat{s} \in \{H, L, \varnothing\}$.

3. $\mu(.|h)$ *is obtained by Bayes' rule if* $\Pr\left(h | m_{L,n}^*, m_{H,n}^*\right) > 0$, *for all* $h \in H^n \cup \{\varnothing, L, H\} \times H^{n-1}$,

4. *The continuation payoff* V *satisfies, for all* $\left(\hat{s}, b; h^{n-1}\right) \in H^n$,

$$V\left(\hat{s}, b; h^{n-1}\right) = E_\mu\left\{\sum_{z=n+1}^{N} \delta^{z-n}\left[\begin{array}{c} u(\theta, b_z^*, \sigma_z) - \Pr(\sigma_z = H)\,\psi_H\left(m_{H,z}^*\right) \\ -\Pr(\sigma_z = L)\,\psi_L\left(m_{L,z}^*\right)\end{array}\right] \middle| \left(\hat{s}, b, h^{n-1}\right)\right\}.$$

I am interested in the PBE of the game when N is large for a fixed $\delta \in [0,1)$. Let $\hat{\theta}_n(h^n)$ denote the Bayes estimator of θ given history h^n

$$\hat{\theta}_n(h^n) \equiv \int \theta dF(\theta | h^n).$$

Note that $F(\theta|h^n)$ is a function of m_H and m_L.

[46] Thus, we write b_n^* instead of $b_n^*\left(\hat{\sigma}_n, h^{n-1}\right)$, and $m_{\sigma,n}^*$ instead of $m_{\sigma,n}^*\left(\sigma, h^{n-1}\right)$.

I assume that $\eta_H > 0$ and that there exists some $\bar{m} > -\eta_L$ with $\psi_L(\bar{m}) \geq \sup_\theta \{u(b,\theta,\sigma)\} - \inf_\theta \{u(b,\theta,\sigma)\}$ for all b, σ.[47] This assumption ensures that the DM never forgets a signal $\sigma_n \in \{H, L\}$ with probability 1.[48] The first issue is whether the Bayes estimator of θ is consistent. In other words, does the DM eventually learn her true attributes after observing a sufficiently large number of signals?

If memory manipulation were constant, the answer would be immediate because in this case, the recollections would be i.i.d., and hence Doob's Consistency theorem would imply that $\hat{\theta}_n(h^n)$ converges to θ. This is formally stated in the following lemma:

Lemma 1 *Suppose* $m_{H,n}\left(h^{n-1}\right) = \tilde{m}_H$ *and* $m_{L,n}\left(h^{n-1}\right) = \tilde{m}_L$ *for all* h^{n-1}, n *and let* $N \to \infty$. *Then* $\hat{\theta}_n \to \theta$ *for almost all histories.*

When memory manipulation is endogenous, however, it is not immediate that the DM eventually learns her true type. Although observed signals σ_n are i.i.d., memory manipulation leads to non-independent and non-identically distributed recollections $\hat{\sigma}_n$. However, because the agent knows the equilibrium strategies, she knows the probability of each signal conditional on the recollection. Therefore, intuitively, the agent correctly updates the recollections and eventually learns her true type regardless of how much manipulation effort she exerts.

The following result will be used in order to show that this intuition is correct:

Lemma 2 *For any fixed history* h^n, $F(\theta|h^n; m_H, m_L)$ *is increasing in* m_H *and decreasing in* m_L.

The lemma above implies that, conditional on reaching each history, the agent always prefers that she had forgotten high signals and remembered low signals. Because the agent is ultimately concerned about σ_n, $F(\theta|h^n; m_H, m_L)$ is not a function of m_H and m_L in all histories that do not contain any $\hat{\sigma}_n = \varnothing$. However, whenever the agent recollects $\hat{\sigma}_n = \varnothing$, she is always better off when she forgets high signals and remembers low signals (since it reduces the probability

[47] This is satisfied, for example, if $\lim_{m_L \to -\eta_L} \psi(m_L) = +\infty$.

[48] Either one of these conditions are needed to ensure identification. If $\eta_H = 0$ and $m_L(h^n) = \eta_L$, then $m_H(h^n) = 0$ for all h^n would imply that $\hat{\sigma}_n = \varnothing$. In this case, the Bayesian posterior would be equal to the prior and, therefore, there is no hope for the Bayes estimator to be consistent. This assumption is not satisfied in the model of Example 1.2.3 ($\eta_H > 0$ is violated). However, it is straightforward to adjust the arguments from this section to establish the same results for that model.

of arriving at $\hat{\sigma}_n = \varnothing$ after a low signal $\sigma_n = L$). Hence, $(\theta|h^n; -\eta_H, 1-\eta_L)$ first-order stochastically dominates $(\theta|h^n; m_H, m_L)$ for all m_H, m_L.[49]

A straightforward implication of Lemma 2 is that:

$$E[\theta|h^n; 1-\eta_H, \bar{m}] \leq E[\theta|h^n; m_H, m_L] \leq E[\theta|h^n; -\eta_H, 1-\eta_L], \tag{1.19}$$

for all m_H and $m_L \geq \bar{m}$ and all histories h^n. But because Lemma 1 implies that both extremes in the inequality (1.19) converge to θ, it thus follows that the term in the middle converges and has limit θ. This result is formally stated in the following proposition:[50]

Proposition 10 (Consistency) *Let $N \to \infty$. Then, $\hat{\theta}_n \to \theta$ for almost all histories.*

Proposition 10 shows that, regardless of the memory manipulation employed by the DM, she eventually learns her true attributes θ. Thus, the benefit of memory manipulation converges to zero, and therefore, memory manipulation converges to zero as the number of observed signals increases:

Proposition 11 (No Manipulation in the Long Run) *Let $N \to \infty$. Then, $m_{H,n} \to 0$ and $m_{L,n} \to 0$ for almost all histories.*

Suppose signals are purely informative. As in Section 1.3, omit the signal σ from the agent's utility function. Define the optimal action $b^O(\theta) \in B$ as the one that maximizes the agent's utility when her attributes θ are known: $b^O(\theta) \in \arg\max_{b \in B} u(b, \theta)$. Proposition 11 implies that $b_n \to b^O$ for almost all histories. Therefore, in the limit, the DM chooses the same actions as an expected utility maximizer who knows θ.

Consider the case of monetary lotteries, and as in Section 1.4, omit the actions from the utility function. The DM's ex-ante utility from observing an additional signal converges to

[49] The first-order dominance (FOSD) is for fixed h^n. Since the probability of each history is itself a function of m_L and m_H, it does not follow that there is unconditional FOSD.

[50] Note that the probability of occurence of a history $\Pr(h^n)$ is a function of the sequence of memory manipulations m_H and m_L. Because $\eta_H > 0$ and $\psi_L(\bar{m}) \geq \sup_\theta \{u(b, \theta, \sigma)\} - \inf_\theta \{u(b, \theta, \sigma)\}$ for some $\bar{m} > -\eta_L$, the sets of histories with zero measuse is the same for all relevant manipulation efforts: $m_H(h^n) \in [-\eta_H, 1-\eta_H]$ and $m_L(h^n) \in [-\eta_L, \bar{m}]$. Therefore, we omit any explicit reference to m_L and m_H when considering almost sure convergence of $\hat{\theta}_n(h^n)$.

$qu(H,\theta)+(1-q)u(L,\theta)$, which is the same utility of an expected utility maximizer when the attributes θ are known.

Therefore, when signals are observed frequently enough, agents will not engage in self-deception and their behavior will converge to the behavior of standard expected utility maximizers. This is consistent with the usual intuition that people do not exhibit ambiguity aversion over frequently observed events or that experts are subject to much less biases (e.g. List, 2003, List and Haigh, 2003).[51]

Remark 6 *In the preceding analysis, we have assumed that A is a singleton so that the DM does not take actions that affect the distribution of the signals σ_a. This assumption simplifies the notation and the proofs. It is not important for our results as long as the first-order stochastic dominance assumption (equation 1.1) is retained. Thus, as long as all signals $a \in A$ are informative about θ, the DM eventually learns her true attributes and does not engage in memory manipulation.*

If the agent has the choice of not observing any signal (see, for example, Subsections 1.6.1 and 1.6.2), then she may choose never to obtain any information about θ. In that case, her expected attributes would not converge.

1.6 Applications

This section presents two applications of the model. The first application provides a self-deception model of the endowment effect. The second application provides a self-deception rationale for people taking sunk investments into consideration when making decisions.

1.6.1 The Endowment Effect

An individual that satisfies the axioms of expected utility theory does not display a difference between the maximum willingness to pay for a good and the minimum compensation demanded to sell the same good (willingness to accept) when income effects are small. However, several

[51] List (2003) also showed that experienced traders of sports paraphernalia show smaller endowment effects for everyday goods used in lab studies than novice traders. This result is also consistent with the model above if the ability to trade sports paraphernalia is correlated with the ability to trade other goods.

empirical works have documented a discrepancy between these values. An individual tends to value one good more when the good becomes part of that person's endowment. Thaler (1980) labeled this phenomenon an "endowment effect."

Kahneman, Knetsch, and Thaler (1990) argued that the endowment effect was caused by loss aversion.[52] This subsection proposes an alternative explanation for the endowment effect. The main idea is that, in most markets, trading requires certain skills or knowledge. At the very least, the parties must form an expectation of how much each good is worth. In more complex markets, they must also estimate the future prices of the goods (which determine the opportunity cost of trading). Therefore, as in the used car example (Subsection 1.2) the outcome from the trade reveals information about how skillful the person is.

As we have seen previously, an individual that cares about her self-image and is subject to imperfect memory will engage in an activity that reveals information about her skills only if the objective value of information is greater than the expected memory cost. Therefore, she may prefer not to trade if the price is only slightly above the expected value of the good.

The model is a special case of the general framework described in Section 1.2. Let $a \in \{T, NT\}$ denote the DM's choice of whether or not to trade an object. Let π denote the gain from trade (in monetary terms), which is unknown by the agent. Trading leads to an outcome $\sigma \in \{H, L\}$, which affects the DM's utility both directly through the gain from trade π_σ and indirectly because it is informative about the DM's skills θ. After observing the outcome σ, the DM engages in memory manipulation $m_L(T)$ and $m_H(T)$. As in Section 1.4, let the DM's preferences over skills θ and money x be represented by $v(\theta) + \tau(x)$ and, with no loss of generality, normalize the monetary payoff from not trading to zero ($\tau(0) = 0$).

Equation (1.13) implies that the DM will prefer to trade if the expected gain from trade is greater than the expected memory cost:

$$q\tau(\pi_H) + (1-q)\tau(\pi_L) \geq q\psi_H(m_H^*(T)) + (1-q)\psi_L(m_L^*(T)).$$

Therefore, we have the following result:

[52] According to loss aversion, losses are weighed substantially more than gains. Then, the cost of losing a good is much higher than the benefit of winning a good.

Proposition 12 (Endowment Effect) *There exist* $\bar{\tau}_1 \geq \bar{\tau}_2 > 0$ *such that:*

1. *the DM agrees to trade in any PBE if* $E[\tau(\pi)] \geq \bar{\tau}_1$,
2. *the DM refuses to trade in any PBE if* $E[\tau(\pi)] \leq \bar{\tau}_2$,
3. *there exist PBE where the DM agrees to trade and PBE where the DM refuses to trade if* $\bar{\tau}_1 > E[\tau(\pi)] > \bar{\tau}_2$.

In particular, a risk neutral individual will demand a strictly positive premium in order to trade. A risk averse individual will demand an even greater premium.

A standard explanation based on ambiguity aversion would argue that the object initially owned by the DM has a less ambiguous distribution than the other object. Therefore, an ambiguity averse agent would not agree to trade if the expected gain from trade is not sufficiently high. Recall from Subsection 1.4.1 that the self-deception model relates the degree of ambiguity aversion with the DM's attributes. Thus, in the present model, the endowment effect is due to the self-evaluation that follows trade. Since the outcome of the trade is informative about the agent's skills or knowledge and therefore leads to costly self-deception, the DM may require some strictly positive premium in order to trade.

1.6.2 Sunk Cost Effects

> The consequences of any single decision (...) can have implications about the utility of previous choices as well as determine future events or outcomes. This means that sunk costs may not be sunk psychologically but may enter into future decisions. (Staw, 1981, pp. 578)

Standard decision theory shows that only incremental costs and benefits should influence decisions. Historical costs, which have already been sunk, should be irrelevant. However, evidence suggests that people often take sunk costs into account when making decisions.[53] Genesove and Mayer (2001), for example, studied the Boston housing market. They have

[53] Sunk costs effects are also called "irrational escalation of commitment", the "entrapment effect", or "too much invested to quit".

shown that when expected prices fall below a the original purchase price, sellers set an asking price that exceeds the asking price of other sellers by between 25 and 35 percent of the difference.

In a field experiment, Arkes and Blumer (1985) randomly selected sixty people to buy season tickets to the Ohio University Theater and divided them in three groups of twenty. Patrons in the first group paid the full price ($15). Those in the second group received a $2 discount, and people in the last group received a $7 discount. Patrons in the first group attended significantly more than those in the discount groups.

This subsection shows that the self-deception model leads to sunk-cost effects. Psychologists have long argued that self-deception may be an important cause of why sunk costs affect choice. For example, Staw (1976) has shown that being personally responsible for an inefficient investment is an important factor in choosing to persist on it. Brockner et al. (1986) have documented that persisting on an inefficient allocation of resources is increased when subjects are told that outcomes reflected their "perceptual abilities and mathematical reasoning."[54]

Whether previous investments succeed or fail has important effects on the decision maker's self-views. Then, as the opening quote suggests, a past choice may be associated with not simply sunk *monetary* costs but also real *psychological* costs. Abandoning a project usually involves admitting that a wrong decision was made. Therefore, an individual revising her position in the project reveals information about her skills or knowledge. As shown in Section 1.3, the DM will prefer to avoid such information if the cost of making an uninformed decision is not high enough. But, in this case, some projects with negative expected value will not be terminated.

The model is a special case of the general framework described in Section 1.2. As in Section 1.4, I assume the DM's utility function is additively separable over attributes θ and money x. For simplicity, I also assume that the DM is risk neutral so that $u(\theta, x) = v(\theta) + x$.

The timing of the model is presented in Figure 1-12. First, the DM chooses whether to invest in a project that costs $K > 0$ and gives a random monetary payoff of π. Let $a_0 \in \{I, NI\}$ denote the investment choice, where $a_0 = I$ if the DM undertakes the investment and $a_0 = NI$ if she does not. After the sunk investment was made, the DM can reevaluate the value of the project at zero cost. Let $a_1 = E$ denote the case where DM reevaluates the project and $a_1 = NE$ otherwise. Reevaluating the project leads to a (purely informative) signal $\sigma \in \{H, L\}$. A high

[54] See Brockner (1992) for a review of the literature.

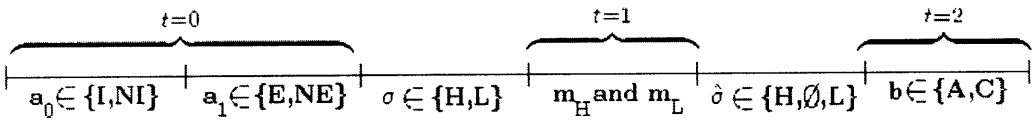

Figure 1-12: Timing of the model

signal is good news, both about the profitability of the project π and about the DM's skills θ.

After observing the signal σ, the DM may engage in memory manipulation m_L and m_H which leads to a recollection $\hat{\sigma} \in \{H, L, \varnothing\}$. Then, she chooses whether or not to abort the project. I write $b = A$ if the project is aborted and $b = C$ if it is continued. If the project is aborted, the DM obtains a monetary payoff of 0. If it is not aborted, the DM has an expected monetary payoff conditional on signal $s \in \{H, L\}$ of π_s.

I assume that the project is ex-ante efficient $E[\pi] > 0$.[55] As was shown in Proposition 6, the agent will prefer to observe the signal σ if the objective value of information, $V = -(1-q)\pi_L > 0$, is greater than the expected manipulation costs, $q\psi_H(m_H^*) + (1-q)\psi_L(m_L^*) > 0$. Hence, if the loss from not aborting after a low signal are "not too large," the DM will prefer not to reevaluate the project:

Proposition 13 (Sunk Cost Effect) *There exist $\bar{\pi}_1 \leq \bar{\pi}_2 < 0$ such that:*

1. *the DM reevaluates the project in any PBE if $\pi_L \leq \bar{\pi}_1$,*
2. *the DM does not reevaluate the project in any PBE if $\pi_L \geq \bar{\pi}_2$, and*
3. *there exist PBE where the DM reevaluates the project and the DM doesn't reevaluate the project if $\bar{\pi}_1 < \pi_L < \bar{\pi}_2$.*

Since reevaluating one's previous decision is informative about the person's skills or knowledge, it leads to self-deception. Therefore, the DM will prefer not to reevaluate her initial choice if the monetary loss π_L from continuing an inefficient project is lower than the expected cost of memory manipulation. Note that the key feature of the model is *not* the psychological cost

[55] If the project is ex-ante inefficient, $E[\pi] \leq 0$, the problem would be trivial since the DM would never invest.

from failure itself. The individual will eventually find out whether the project is successful or not. However, by not reevaluating a project, the individual avoids the psychological cost from *self-deception*.[56]

1.7 Conclusion

This chapter proposed a model of choice under risk based on imperfect memory and self-deception. The model provides a unified explanation for a number of biases in decision-making. It also leads to non-expected utility representation that is consistent with recent experimental evidence relating ambiguity aversion to an individual's skills or knowledge.

The model can be enriched in several directions by incorporating strategic components. Principal-agent relationships seem like a natural application of the theory. Since the outcome of the relationship is typically informative about the agent's skill or knowledge, principals may prefer to offer contracts that do not completely reveal the outcome to the agent. Therefore, firms may prefer not to condition wages on economy-wide shocks. Similarly, CEOs may be "rewarded for luck."

Another interesting direction is in the field of incomplete contracts. Contracts may be incomplete due to the contracting parties' preferences for avoiding information correlated with their skills or knowledge.[57] However, because parties understand the consequences of contract forms and post-contractual decisions, the allocation of rights may matter for the outcomes. Therefore, the general framework proposed here may provide a behavioral model for a theory of ownership based on incomplete contracts.

The model can also be embedded in a general equilibrium model. Since self-deception leads to endowment effects, the model may provide an explanation for the low volume of trades of uncertain assets occurring in equilibrium.[58]

Finally, the model can lead to interesting predictions when θ is interpreted as a parameter of

[56] The argument above is related to agency explanations. For example, as argued by Li (2007), in environments with adverse selection, agents may prefer not to change their opinions if this publicly conceals bad news about their abilities. It is unclear, however, whether agency concerns would play an important role in contexts where the decisions are not publicly observed.

[57] Mukerji (1998) showed that ambiguity aversion may lead to incomplete contracts Tirole (2008) considered a model where thinking about contingencies is costly. In his model, contracts may be "too complete."

[58] See Billot et al. (2000) for a model based on ambiguity aversion.

anticipatory utility. Because anticipatory utility typically leads to a first-order gain from memory manipulation but only second-order costs through suboptimal decision-making, individuals will forget negative news and remember positive news with probability above their natural rates. For example, a model of portfolio allocation where signals σ are informative about the profitability of a risky asset may provide an explanation for why most investors hold extremely underdiversified portfolios and overinvest in stocks issued by the their employing firm.

Appendix A Non-Separable Preferences

In Section 1.4, the DM's preferences were assumed to be additively separable between attributes and money. In this section, I consider general utility functions. It turns out that a main feature in this general model is the degree of complementarity between attributes and money. As will be discussed later, since a DM is not as affected by monetary outcomes when she is uninformed about her attributes when attributes and money are complementary, complementarity can be interpreted as providing "psychological insurance." Therefore, the DM may prefer a lottery whose outcomes are informative about her attributes if the complementarity effect is greater than the costs of self-deception. Moreover, the resulting probability weighting function may have an "inverted S-shape" as in Tversky and Kahneman (1992) and Prelec (1998).

Let $u_\sigma(s) \equiv \int u(\theta, s)\, dF(\theta|\sigma)$ denote the expected utility from a monetary amount equal to s conditional on outcome σ. As in Corollary 2, it can be shown that, in any PBE, $m_H^* > 0 \geq m_L^*$.[59]

Define the degree of complementarity between θ and money by

$$\chi(H, L) \equiv u_H(H) + u_L(L) - u_L(H) - u_H(L). \tag{1.20}$$

Note that $\chi(H, L) \geq 0$ if u has increasing differences and $\chi(H, L) \leq 0$ if u has decreasing differences. The additively separable case presented in the text corresponds to the case where $\chi(H, L) = 0$. The ex-ante expected utility from the lottery is

$$\begin{aligned} U(\Sigma) &= q(\eta_H + m_H^*)\, u_H(H) + (1-q)(\eta_L + m_L^*)\, u_L(L) \\ &\quad + q(1 - \eta_H - m_H^*)\, u_\varnothing(H) + (1-q)(1 - \eta_L - m_L^*)\, u_\varnothing(L) - MC, \end{aligned}$$

where $MC = q\psi_H(m_H^*) + (1-q)\psi_H(m_L^*)$ is the expected memory cost. Then, long but tedious

[59] As in Corollary 2, it can also be shown that

$$\begin{aligned} u_H(H) - u_L(L) &\geq \psi_H'(1-\eta_H) \Longrightarrow m_H^* = 1 - \eta_H,\ m_L^* = 0, \text{ and} \\ u_H(H) - u_L(L) &< \psi_H'(1-\eta_H) \Longrightarrow 0 < m_H^* < 1 - \eta_H,\ m_L^* < 0. \end{aligned}$$

algebraic manipulations yield

$$U(\Sigma) = qu_H(H) + (1-q)u_L(L) + z\chi(H,L) - MC. \tag{1.21}$$

where $z = \frac{q(1-q)(1-\eta_L - m_L^*)(1-\eta_H - m_H^*)}{q(1-\eta_H - m_H^*)+(1-q)(1-\eta_L - m_L^*)} > 0$ and $MC = q\psi_H(m_H^*) + (1-q)\psi_H(m_L^*)$.

The utility of a monetary lottery can be decomposed in three terms: First, the expected utility $qu_H(H) + (1-q)u_L(L)$ of the lottery when memory is perfect. Second, the expected manipulation costs MC. These two effects are precisely the same as in the additively separable case (see equation 1.13). The third effect, which is not present when the utility is additively separable, is the degree of complementarity between attributes and money. When signals are forgotten, there is probability $\alpha(m_H^*, m_L^*)$ that a high signal was observed and the complementary probability that a low signal was observed. Thus, forgetting a signal can be seen as providing "psychological insurance" to the agent. This raises her expected utility if θ and money are complementary ($\chi > 0$) and decreases her expected utility if they are substitutes ($\chi < 0$).

Proceeding as in Subsection 1.4.1, it follows that the DM's expected utility can be represented by

$$U(\Sigma) = w(q) \times u_H(H) + [1 - w(q)] \times u_L(L),$$

where $w(q) = q + \frac{z\chi(H,L) - MC}{u_H(H) - u_L(L)}$. Moreover, it is straightforward to show that $w(0) = 0$, and $w(1) = 1$. Therefore, when attributes and money are complementary, the DM may exhibit ambiguity loving behavior. In particular, the following example shows that the model may lead to an inverted S-shaped probability weighting function:

Example 7 (Inverted S-shaped Probability Weighting Function) *Consider the limited memory model of Example 2 and suppose that the manipulation effort is a binary variable:* $m_H \in \{0, \frac{3}{4}\}$, *where* $\psi_H(\frac{3}{4}) = \frac{1}{5}$. *Let* $\chi(H,L) = u_H(H) - u_H(L) = 1$. *Then, self 1 chooses to engage in memory manipulation if* $q \in (0, \frac{11}{12})$. *It is straightforward to show that, for values of* q *such that the DM engages in memory manipulation, the probability weighting function has an inverted S-shape:*

$$w(q) \begin{cases} > q \text{ if } q \in (0, \frac{1}{2}) \\ < q \text{ if } q \in (\frac{1}{2}, \frac{11}{12}) \end{cases}.$$

As in Section 1.4, denote by U^I the utility of a lottery with the same distribution over monetary outcomes as the one above but whose monetary outcomes are uninformative about θ. Rearranging equation (1.21), we obtain

$$U(\Sigma) = U^I + y\chi(H, L) - MC, \tag{1.22}$$

where $y = q(1-q)\left[1 + \frac{(1-\eta_L - m_L^*)(1-\eta_H - m_H^*)}{q(1-\eta_H - m_H^*)+(1-q)(1-\eta_L - m_L^*)}\right] > 0$. Consider the choice between the lottery Σ and another lottery with the same distribution over monetary outcomes but whose monetary outcomes are uninformative about θ. Equation (1.22) implies that the DM will prefer lottery Σ if the degree of complementarity is high enough or if the expected memory cost is low enough: $y\chi(H, L) \geq MC$. Therefore, when attributes and money are complementary, the DM may prefer the attribute-dependent lottery.

However, when the monetary lottery is "small" (i.e., when the lottery pays $x = \varepsilon s$ for ε low), the complementarity effect vanishes. Since the memory cost converges to a strictly positive number as ε converges to zero, it follows that the certainty equivalent of the lottery converges to $CE(0) < 0$. Therefore,

$$\lim_{\varepsilon \to 0_+} \frac{\pi(\varepsilon)}{\varepsilon} = -\lim_{\varepsilon \to 0_+} \frac{CE(\varepsilon)}{\varepsilon} = +\infty$$

and, for any degree of complementarity between attributes and money, the DM always exhibits zeroth-order risk aversion. This is formally stated in the following proposition:

Proposition 14 *In any PBE, the DM exhibits zeroth-order risk aversion.*

It is interesting to contrast the general model with the a model from the following example where the DM does not face memory costs:

Example 8 (Exogenous Memory Model) *Take $\psi_s(m_s) = +\infty$ for all $m_s \neq 0$. Thus, the agent cannot engage in endogenous memory manipulation. Let $\eta_s < 1$ so that the agent forgets outcome with (exogenous) probabilities $1 - \eta_s > 0$. If $\eta_H > \eta_L$, memory is selective in the sense that good news is more likely to be remembered than bad news.*

When memory manipulation is endogenous (and differentiable at $m_s = 0$), the effect from memory manipulation always dominates the complementarity effects and the DM displays

zeroth-order risk aversion. When memory manipulation is exogenous, the order of risk aversion is determined by the degree of complementarity between attributes and money. Note that for small ε, attributes and money are complementary if $u'_H(0) < u'_L(0)$ and substitutes if $u'_H(0) > u'_L(0)$.

Proposition 15 *In the exogenous memory model: (i) the DM is first-order risk averse if* $u'_L(0) > u'_H(0)$; *(ii) the DM is first-order risk seeking if* $u'_L(0) < u'_H(0)$; *and (iii) the DM has second-order risk preferences if* $u'_L(0) = u'_H(0)$.

Therefore, the DM may display risk preferences of first order when there are no manipulation costs. Unlike when memory manipulation is endogenous, the DM may be first-order risk seeking or have risk preferences of second order.

Appendix B Non-Bayesian Framework

Throughout the chapter, I have maintained the assumption that the DM understands that she engages in memory manipulation and, thus, interprets her recollections according to Bayes' rule. Therefore, in the model presented in the text, individuals are sophisticated. In this section, I consider the case of naive individuals. As in Mullainathan (2002), naive individuals are unaware of their imperfect memory and interpret recollections as if they were the true outcomes. Two interesting features arise under naiveté. First, unlike the model of sophisticated individuals, the equilibrium is unique. Second, decision makers may prefer to observe a signal even if it has no objective value. As a consequence, they may display ambiguity seeking behavior even under additive separability between attributes and money. Moreover, the individual may exhibit zeroth-order risk seeking behavior.

Consider a naive decision maker (NDM), who is unaware of her memory manipulation efforts. Therefore, she applies Bayes' rule as if her recollections were generated by the memory system when she does not engage in memory manipulation, i.e. $m_L = m_H = 0$. When $\hat{\sigma} \in \{H, L\}$, she correctly infers that outcome $\sigma = \hat{\sigma}$ has been observed in period 1. However, when an outcome

is forgotten, she attributes weight

$$\rho \equiv \frac{q(1-\eta_H)}{q(1-\eta_H)+(1-q)(1-\eta_L)} \tag{1.23}$$

to a high outcome and $(1-\rho)$ to a low outcome. I refer to such updating rule as *naive Bayes' rule.* The following definition proposes an adaptation of the PBE concept to naive decision makers:

Definition 3 *A Perfect Naively Bayesian Equilibrium (PNBE) of the game is a strategy profile* $(a^*, b^*, m_H^*(a), m_L^*(a))$ *and posterior beliefs* $\mu(.|\hat{\sigma}_a)$ *such that:*

1. $a^* \in \arg\max_{a \in A} \left\{ \begin{array}{c} E_{\hat{\sigma}_a}\left[E_\mu\left[u\left(a, b_a^*(\hat{\sigma}_a), \theta, \sigma_a\right)|\hat{\sigma}_a\right]|m_L^*(a), m_H^*(a)\right] \\ -q\psi_H\left(m_H^*(a)\right) - (1-q)\psi_L\left(m_L^*(a)\right) \end{array} \right\}$;

2. $m_s^*(a) \in \arg\max_{m_s} \left\{ \begin{array}{c} (\eta_s + m_s) E_\mu\left[u\left(a, b_a^*(\hat{\sigma}_a), \theta, s\right)|\hat{\sigma}_a = s\right] \\ +(1-\eta_s-m_s) E_\mu\left[u\left(a, b_a^*(\hat{\sigma}_a), \theta, s\right)|\hat{\sigma}_a = \varnothing\right] - \psi_s(m_s) \end{array} \right\}$,
   $s \in \{H, L\}$;

3. $b_a^*(\hat{\sigma}) \in \arg\max_{b \in B} \left\{E_\mu\left[u(a, b, \theta, \sigma_a)|\hat{\sigma}_a = \hat{\sigma}\right]\right\}$;

4. $\mu(\theta|\hat{\sigma}_a = \hat{\sigma})$ *is obtained by naive Bayes' rule if* $\Pr(\hat{\sigma}_a = \hat{\sigma}|m_L = m_H = 0) > 0$, $\forall \hat{\sigma} \in \{L, H, \varnothing\}$.

Conditions $1-3$ are the same as in the PBE concept. Condition 4 modifies the standard Bayesian condition by requiring agents to follow the naive Bayes rule instead.

An interesting special case of this naive framework is obtained when we take the forgetfulness memory system of Example 1. Recall that if the state $\varnothing$ is interpreted as a recollection of a high outcome, then the model from Example 1 becomes one where the agent is able to convince herself that a low outcome was a high outcome by engaging in memory manipulation. Suppose an individual recollects a high outcome (i.e., $\hat{\sigma} = \varnothing$). If this individual is sophisticated, she then corrects for her memory imperfection and attributes some (Bayesian) weight to the possibility that she has observed a low outcome but managed to convince herself that the outcome was high instead. On the other hand, a naive individual believes her recollection is correct and attributes full weight to a high outcome ($\rho = 1$).

B1 Equilibrium Uniqueness

This subsection establishes that a PNBE exists and, under mild conditions, is unique. The naive updating rule implies that the NDM's expected utility given $\hat{\sigma} = \varnothing$ is

$$u_{\varnothing}(a, b, \sigma) = \rho u_H(a, b, \sigma) + (1 - \rho) u_L(a, b, \sigma). \tag{1.24}$$

Upon observing an outcome $s \in \{H, L\}$, self 1 maximizes:

$$(\eta_s + m_s) u_s(a, b_a(s), s) + (1 - \eta_s - m_s) u_{\varnothing}(a, b_a(\varnothing), s) - \psi_s(m_s). \tag{1.25}$$

The key feature of the naive updating rule is that it is not a function of the amount of memory manipulation employed by self 1. This greatly simplifies the computation of the PNBE of the model since, unlike in the sophisticated case, there is no feedback between self 2's expectation of the manipulation exerted by self 1 and self 1's manipulation choice. Then, the equilibrium amount of manipulation is determined by the maximum of expression (1.25).

Proposition 16 *There exists a PNBE. Furthermore, if ψ_s is strictly convex and $b_a(\hat{s})$ is a (single-valued) function where $s \in \{H, L\}$ and $\hat{s} \in \{H, L, \varnothing\}$, the PNBE is essentially unique.*[60]

Proof. Existence follows the same argument as Proposition 1. Note that Condition 3 from Definition 3 implies that $b_a(\hat{s})$ is not a function of self 1's memory manipulation. Strict convexity of ψ_s implies that expression (1.25) is strictly concave in m_s. Then, the equilibrium amounts of memory manipulation m_L^* and m_H^* are unique. Condition 4 implies that beliefs must also coincide in all recollections such that $\Pr(\hat{\sigma}_a = \hat{\sigma} | m_L^* = m_H^* = 0) > 0$. ■

Corollary 3 *The PNBE is essentially unique when either: (i) $u(a, ., \theta, \sigma_a) : B \to \mathbb{R}$ is a strictly concave function, or (ii) B is a singleton (i.e., the individual does not take ex-post actions).*

[60] *The PNBE is essentially unique in the sense that, all PNBE feature the same choices of actions a and b, manipulation efforts m_L and m_H, and beliefs given recollections that are believed to be reached with positive probability (i.e., $(\hat{\sigma}_a = \hat{\sigma} | m_L = m_H = 0) > 0$). Equilibria may diverge only with respect to beliefs at recollections that are not believed to be reached with positive probability. Obviously, one can ensure uniqueness of beliefs in all recollections by assuming that the NDM believes that all recollections are reached with positive probability:* $0 < \min\{\eta_H, \eta_L\} < 1$.

Remark 7 *Suppose that B is finite and fix the natural rates of remembering an outcome η_L and η_H. Since the set of utility functions $u : \Theta \times A \times B \times \mathbb{R} \to \mathbb{R}$ such that $\arg\max_{b \in B} \{E_\mu [u(a, b, \theta, \sigma_a) | \hat{\sigma}_a = \hat{\sigma}]\}$ contains more than one element is nowhere dense, it follows that the PNBE is essentially unique for generic utility functions when the set of ex-post actions B is finite.*

The generic uniqueness of the PNBE contrasts with the multiplicity of the PBE discussed in Subsubsection 1.2.3. Multiplicity arises from the fact that self 1 affects self 2's equilibrium inference when the individual is sophisticated. In the naive case, because there is not effect from memory manipulation on self 2's inference, uniqueness is obtained.

B2 Ambiguity-Seeking Behavior

For simplicity, consider the forgetfulness memory system of Example 1 and, as in the Section 1.4, assume that the utility is additively separable between attributes and money. Then, the equilibrium amount of memory manipulation is $m_L^* = \min\left\{\psi_L'^{-1}(\Delta u); 1\right\}$. The ex-ante expected utility of the NDM is

$$\begin{aligned} U(\Sigma) &= (1-q)(1+m_L^*)u_L + [q-(1-q)m_L^*]u_\varnothing - (1-q)\psi_L(m_L^*) \\ &= (1-q)(1+m_L^*)u_L + [q-(1-q)m_L^*]u_H - (1-q)\psi_L(m_L^*), \end{aligned}$$

where the second inequality uses the fact that $u_\varnothing = u_H$. The NDM prefers to observe the signal Σ if and only if the expected improvement in self-image $|m_L^*|\Delta u$ is greater than the cost of memory manipulation $\psi_L(m_L^*)$.[61] Thus, naive individuals may prefer to observe signals even if the objective value of information (which in this case is zero) is lower than the expected costs of manipulation.

Remark 8 *Proceeding as in Proposition 8, it follows that the NDM's expected utility from the monetary lottery can be represented by*

$$U(\Sigma) = w(q)u_H + [1-w(q)]u_L,$$

[61] As in Subsection 1.3.1, the NDM's surplus from observing a signal is decreasing in the favorableness of her prior distribution over her attributes under Assumption 3. However, unlike Conjecture 1, this surplus may be positive when the individual is naive.

where $w(q) = q - (1-q)\left[m_L^* + \frac{\psi_L(m_L^*)}{\Delta u}\right]$, $w(0) = 0$, *and* $w(1) = 1$. *Thus, the NDM is ambiguity averse if* $|m_L^*|\,\Delta u < \psi_L(m_L^*)$ *and ambiguity seeking if the reverse inequality is satisfied. Hence, a naive individual may be ambiguity seeking even when the utility function is additively separable between attributes and money.*

B3 Zeroth-Order Risk Seeking Behavior

This subsection shows that the NDM may be zeroth-order risk seeking. As in Subsection 1.4.3, consider a lottery that pays $x = \varepsilon s$, $s \in \{H, L\}$, where $qH + (1-q)L = 0$. Let $m_s^*(\varepsilon)$ denote the equilibrium amount of memory manipulation as a function of ε. The certainty equivalent of this lottery is:

$$\begin{aligned}\int u(\theta, CE(\varepsilon))\,dF(\theta) &= (1-q)(1+m_L^*(\varepsilon))\,u_L(L) \\ &\quad + [q-(1-q)\,m_L^*(\varepsilon)]\,u_H(H) - (1-q)\,\psi_L(m_L^*(\varepsilon)) - q\psi_H(m_H^*(\varepsilon)),\end{aligned}$$

Recall that $u_\sigma(s) \equiv \int u(\theta, s)\,dF(\theta|\sigma)\,u_s(\sigma) = v_\sigma + \tau(s)$, where the last equality follows from additive separability. Then, taking the limit as $\varepsilon \to 0_+$, we obtain:

$$\tau(CE(0)) = -m_L^*(1-q)\,\Delta v - (1-q)\,\psi_L(m_L^*) - q\psi_H(m_H^*).$$

Hence, $CE(0) > 0$ if $|m_L^*(0)|\,(v_H - v_L) > \psi_L(m_L^*(0)) + \frac{q}{1-q}\psi_H(m_L^*(0))$ and the NDM is zeroth-order risk seeking. In the opposite case, the NDM is zeroth-order risk averse. Thus, we have established the following result:

Proposition 17 *The NDM is:*

- *zeroth-order risk averse if* $|m_L^*(0)|\,(v_H - v_L) < \psi_L(m_L^*(0)) + \frac{q}{1-q}\psi_H(m_L^*(0))$, *and*
- *zeroth-order risk seeking if* $|m_L^*(0)|\,(v_H - v_L) > \psi_L(m_L^*(0)) + \frac{q}{1-q}\psi_H(m_L^*(0))$.

Appendix C Finite Number of Realizations

In the main text, we assume that each outcome σ_a may be either high or low. It is straightforward to generalize this framework to allow for any finite number of possible outcomes. Suppose that, given action $a \in A$, an outcome $\sigma_a \in \{1, 2, ..., S_a\}$ is realized, $S_a \geq 2$. Outcomes are ordered by first-order stochastic dominance:

$$F(\theta|\sigma_a = s) \leq F(\theta|\sigma_a = s+1)$$

for all $\theta \in \Theta$, $s \in \{1, 2, ..., S_a\}$ and $a \in A$, with strict inequality for some value of θ.

An outcome $s \in \{1, 2, ..., S_a\}$ is remembered with probability $\eta_{s,a} + m_s$, where $\eta_{s,a} \in [0, 1]$. Self 1 exerts memory manipulation $m_s \in [-\eta_{s,a}, 1 - \eta_{s,a}]$, which costs $\psi_s(m_s) \geq 0$. Then, self 2 observes a recollection of the outcome σ_a, which is denoted by $\hat{\sigma}_a \in \{1, 2, ..., S_a, \varnothing\}$ and takes an action $b \in B$.

Preferences are represented by a von Neumann-Morgenstern utility function $u : \Theta \times A \times B \times \mathbb{R} \to \mathbb{R}$ which is strictly increasing in θ. When $u(\theta, a, b, x) = u(\theta, a, b, y)$ for all $x, y \in \mathbb{R}$, the model has *purely informative signals.* If signals are purely informative and A and B are singletons, we say that they have a *purely hedonic value.* When $u(\theta, a, b, x) \neq u(\theta, a, b, y)$ for some $x, y \in \mathbb{R}$ we say that the model has *monetary signals.*

It is straightforward to generalize the results in the text to this framework. For the representation result of Proposition 8, however, one should note that probability weights are no longer unique when $S_a > 2$.

Entrepreneurship Example The performance $s \in \{S, F\}$ of an entrepreneur is affected by two independent variables: her attributes θ and the external conditions $r \in \{1, 2, 3, ..., R\}$. Attributes and external conditions are substitutes for the entrepreneur's performance. Therefore, given her performance $s \in \{S, F\}$, more favorable external conditions r provide bad news about the agent's attributes (in the sense of first-order stochastic dominance). The individual always recollects her performance s, but may manipulate her memory in order to change the rate at which she remembers the external conditions r.

This situation is modeled as follows. Let $\sigma \in \{S, F\} \times \{1, ..., R\}$ denote the outcome of the

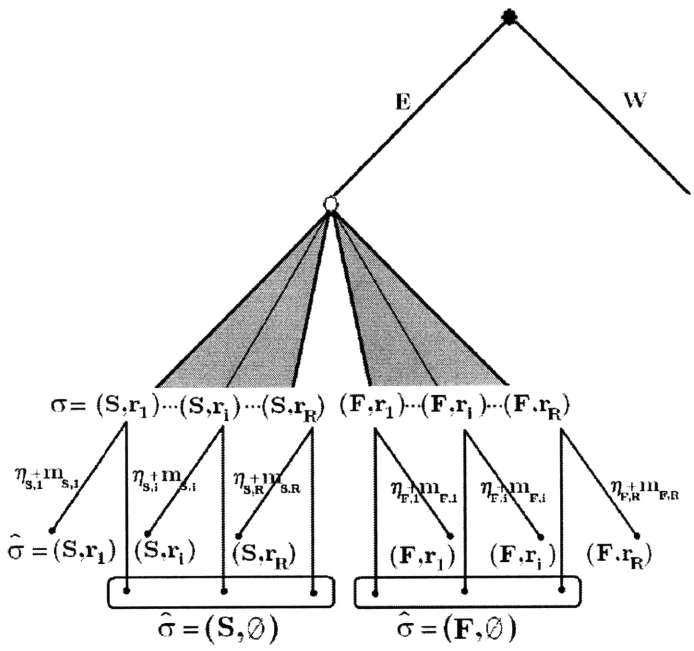

Figure 1-13: Entrepreneur Example

project. Outcomes are ordered by first-order stochastic dominance:

$$(S,1) \succeq_{FOSD} (S,2) \succeq_{FOSD} ... \succeq_{FOSD} (S,R) \succeq_{FOSD} (F,1) \succeq_{FOSD} ... \succeq_{FOSD} (F,R),$$

where we write $x \succeq_{FOSD} y$ if x first-order stochastically dominates y. Given an outcome (s,r), self 1 chooses the probability at which the external conditions r are forgotten by exerting manipulation effort $m_{s,r}$. Then, self 2 applies Bayes' rule to the recollections $(s,\hat{r})$, where $\hat{r} \in \{r, \varnothing\}$. The agent's payoff net of manipulation costs given a recollection $(s,\hat{r})$ is

$$E[v(\theta,s)|s,\hat{r}] + \tau(s),$$

where $s \in \{S,F\}$ and $\hat{r} \in \{1,2,...,R,\varnothing\}$. Figure 1-13 presents the agent's decision tree.

It is straightforward to extend the results from the general framework to this environment. In particular, expected manipulation costs are always strictly positive. Therefore, the agent will require the expected monetary payoffs from starting a new company to be strictly higher than the payoff from the previous job in order to become an entrepreneur. Moreover, if all outcomes

have the same natural rate of recollection (i.e., $\eta_{s,r} = \eta$ for all s, r), then $m_{s,1} \geq m_{s,2} \geq ... \geq m_{s,R}$ with at least one strict inequality, $s \in \{S, F\}$. Hence, the agent will remember negative external conditions more frequently than positive ones.

Appendix D Proofs

D1 Proofs of Propositions and Lemmas

Proof of Proposition 1: Define $\hat{m}_L(m_L, m_H)$ and $\hat{m}_H(m_L, m_H)$ as the set of maxima of (1.4) and (1.5), respectively (these are the best-response correspondences of self 1). Since (1.4) and (1.5) are continuous and concave functions defined over a compact set, $\hat{m}_L(m_L, m_H)$ and $\hat{m}_H(m_L, m_H)$ are non-empty, convex, and compact sets. Define the transformation $T : [-\eta_L, 1-\eta_L] \times [-\eta_H, 1-\eta_H] \to [-\eta_L, 1-\eta_L] \times [-\eta_H, 1-\eta_H]$ by

$$T(m_L, m_H) = (\hat{m}_L(m_L, m_H), \hat{m}_H(m_L, m_H)).$$

Then, Kakutani's theorem establishes that there exists a fixed-point of T which is a PBE. ∎

Proof of Proposition 2: The proof will use the following result:

Claim A1. $u_H(a, b_a(\varnothing), L) > u_L(a, b_a(L), L)$.

Proof of the claim. By revealed preference, $u_\varnothing(a, b_a(\varnothing), L) \geq u_\varnothing(a, b_a(L), L)$. From the definition of $u_\varnothing$,

$$\alpha u_H(a, b(\varnothing), L) + (1-\alpha) u_L(a, b(\varnothing), L) \geq \alpha u_H(a, b(L), L) + (1-\alpha) u_L(a, b(L), L),$$

where I omit the efforts m_H and m_L from $\alpha(m_H, m_L)$ for notational clarity. Rearranging, gives

$$\alpha [u_H(a, b_a(\varnothing), L) - u_H(a, b_a(L), L)] \geq (1-\alpha) [u_L(a, b_a(L), L) - u_L(a, b_a(\varnothing), L)].$$

Then, revealed preference implies that

$$u_H(a, b_a(\varnothing), L) - u_H(a, b_a(L), L) \geq \frac{1-\alpha}{\alpha} [u_L(a, b_a(L), L) - u_L(a, b_a(\varnothing), L)] \geq 0.$$

Thus, $u_H(a, b_a(\varnothing), L) \geq u_H(a, b_a(L), L)$. But, first-order stochastic dominance implies that $u_H(a, b_a(L), L) >$

$u_L(a, b_a(L), L)$. Therefore, we have that $u_H(a, b_a(\varnothing), L) > u_L(a, b_a(L), L)$. ∎

Proof of Proposition 2: Because $\eta_H < 1$, the set of strictly positive efforts given a high signal $(0, 1-\eta_H]$ is nonempty. Since (1.5) is strictly concave, it suffices to show that its derivative evaluated at $m_H = 0$ is strictly positive:

$$u_H(a, b(H), H) - u_L(a, b(\varnothing), H) - \alpha(m_L^*(a), m_H^*(a))[u_H(a, b(\varnothing), H) - u_L(a, b(\varnothing), H)] > 0, \quad (1.26)$$

for all $m_L^*(a), m_H^*(a)$, where I have used the fact that $\psi_H'(0) = 0$.

Note that, by revealed preference, $u_H(a, b(H), H) \geq u_H(a, b(\varnothing), H)$. Hence,

$$\frac{u_H(a, b(H)) - u_L(a, b(\varnothing))}{u_H(a, b(\varnothing)) - u_L(a, b(\varnothing))} \geq 1.$$

Rearranging, we obtain:

$$\begin{aligned} & u_H(a, b(H), H) - u_L(a, b(\varnothing), H) - \alpha(m_L^*(a), m_H^*(a))[u_H(a, b(\varnothing), H) - u_L(a, b(\varnothing), H)] \\ \geq\ & u_H(a, b(H), H) - u_L(a, b(\varnothing), H) - [u_H(a, b(\varnothing), H) - u_L(a, b(\varnothing), H)] \geq 0. \end{aligned}$$

This shows that the expression on the left-hand side of (1.26) is non-negative. Suppose it is equal to zero. Then, by the previous inequality, it must be the case that $\alpha(m_L^*(a), m_H^*(a)) = 1$. But $\alpha(m_L^*(a), m_H^*(a)) = 1$ implies that $m_L^* = 1 - \eta_L$ which, from the Kuhn-Tucker condition of the maximization of (1.4), requires that

$$u_L(a, b_a(L), L) - u_H(a, b_a(\varnothing), L) \geq \psi_L'(1 - \eta_L) \geq 0.$$

But, from Claim A1, $u_L(a, b_a(L), L) - u_H(a, b_a(\varnothing), L) < 0$, which contradicts the inequality above. Hence, $m_H^*(a) > 0$ for all $a \in A$. ∎

Proof of Proposition 3: Define the function W_s as the expected utility of self 1 conditional on $\sigma = s$:

$$\begin{aligned} W_s(m_L, m_H, a, \{b_a\}) = & (1 - \eta_s - m_s)[\alpha(m_L, m_H) u_H(a, b_a(\varnothing), H) + [1 - \alpha(m_L, m_H)] u_L(a, b_a(\varnothing), L)] \\ & + (\eta_s + m_s) u_s(a, b_a(s), s) - \psi_s(m_s) \end{aligned}.$$

For notational clarity, I omit the term a from $m_s^*(a)$. In any PBE, m_s^* solves:

$$\max_{m_s} \left\{ \begin{array}{c} (1-\eta_s-m_s)\left[\alpha\left(m_L^*, m_H^*\right) u_H\left(a, b_a(\varnothing), H\right) + \left[1-\alpha\left(m_L^*, m_H^*\right)\right] u_L\left(a, b_a(\varnothing), L\right)\right] \\ +\left(\eta_s+m_s\right) u_s\left(a, b_a(s), s\right) - \psi_s\left(m_s\right) \end{array} \right\}.$$

Therefore, the envelope theorem implies that

$$\left.\frac{\partial W_s}{\partial m_L}\right|_{\substack{m_L=m_L^* \\ m_H=m_H^*}} = (1-\eta_s-m_s^*)\left[u_H\left(a, b_a(\varnothing), H\right) - u_L\left(a, b_a(\varnothing), L\right)\right] \frac{\partial \alpha\left(m_L^*, m_H^*\right)}{\partial m_s}. \tag{1.27}$$

The DM's ex-ante utility is equal to

$$\mathcal{U}\left(m_H, m_L, a, \{b_a\}\right) = q W_H\left(m_L, m_H, a, \{b_a\}\right) + (1-q) W_L\left(m_L, m_H, a, \{b_a\}\right)$$

Thus,

$$\left.\frac{\partial \mathcal{U}\left(m_H, m_L, a, \{b_a\}\right)}{\partial m_s}\right|_{\substack{m_H=m_H^* \\ m_L=m_L^*}} = q \frac{\partial W_H\left(m_H^*(a), m_L^*(a), a\right)}{\partial m_s} + (1-q) \frac{\partial W_L\left(m_H^*(a), m_L^*(a), a\right)}{\partial m_s}.$$

Since $\frac{\partial \alpha\left(m_H, m_L\right)}{\partial m_H} \leq 0 \leq \frac{\partial \alpha\left(m_H, m_L\right)}{\partial m_L}$ with at least one inequality being strict, equation (1.27) yields

$$\frac{\partial \mathcal{U}\left(m_H, m_L, a, \{b_a\}\right)}{\partial m_H} < 0 < \frac{\partial \mathcal{U}\left(m_H, m_L, a, \{b_a\}\right)}{\partial m_L}.$$

Then, the result follows from the concavity of $\mathcal{U}$. ∎

Proof of Proposition 4: From Proposition 2, it follows that $m_H^*(a) > 0$. First, we establish that $m_L^*(a) \leq 0$. By the strict concavity of equation (1.4), it suffices to show that its derivative evaluated at $m_L = 0$ is weakly negative:

$$-\alpha\left(m_L^*(a), m_H^*(a)\right) \Delta u - \underbrace{\psi_L'(0)}_{0} \leq 0,$$

which is true because $\alpha\left(m_L^*(a), m_H^*(a)\right) \geq 0$ and $\Delta u > 0$.

Let $\Delta u > \psi_H'(1-\eta_H)$. Suppose, in order to obtain a contradiction that there exists a PBE with $m_L^* = 0$. Then, from Kuhn-Tucker's conditions of the maximization of (1.4),

$$\alpha\left(0, m_H^*(a)\right) \Delta u = 0$$

for some $m_H^*(a)$ that maximizes (1.5) given $m_L^*(a) = 0$. Because $\Delta u > 0$, this is satisfied if and only if $\alpha\left(0, m_H^*(a)\right) = 0$. But $\alpha\left(0, m_H^*(a)\right) = 0$ implies that $m_H^* = 1 - \eta_H$. From Kuhn-Tucker's conditions of

the maximization of (1.5), there exists a PBE with $m_H^* = 1 - \eta_H$ if and only if

$$[1 - \alpha(m_L^*(a), m_H^*(a))]\Delta u \geq \psi_H'(1 - \eta_H),$$

for some m_L^* that maximizes (1.4). Substituting $\alpha(m_L^*(a), m_H^*(a)) = 0$, it follows that $\Delta u \geq \psi_H'(1 - \eta_H)$, which contradicts $\Delta u > \psi_H'(1 - \eta_H)$. ∎

Proof of Proposition 5: Existence follows from Proposition 2. For a fixed m_L^*, self 1 solves:

$$\max_{-1 \leq m_L \leq 0} u_L - m_L \alpha(m_L^*, 0)\Delta u - \psi_L(m_L).$$

The Kuhn-Tucker conditions are:

$$\begin{aligned} \alpha(m_L^*, 0)\Delta u &\geq -\psi_L'(-1) \implies m_L = -1, \\ \alpha(m_L^*, 0)\Delta u &\leq 0 \implies m_L = 0, \text{ and} \end{aligned}$$

$$0 < \alpha(m_L^*, 0)\Delta u < -\psi_L'(-1) \implies \alpha(m_L^*, 0)\Delta u = -\psi_L'(m_L).$$

In the PBE, $m_L = m_L^*$. Substituting $\alpha(m_L, 0) = \frac{q}{q-(1-q)m_L}$ and using the implicit function theorem, it follows that the unique PBE has manipulation efforts:

$$\begin{aligned} m_L^* &= -1 \text{ if } \Delta u \geq -\frac{\psi_L'(-1)}{q}, \text{ and} \\ \frac{q}{q-(1-q)m_L^*}\Delta u &= -\psi_L'(m_L^*) \text{ if } \Delta u < -\frac{\psi_L'(-1)}{q}. \end{aligned}$$

The first claim follows by inspection. Let the cost of manipulation be $\psi_L(m_L, \kappa)$, where κ parametrizes the marginal cost of memory manipulation: $\frac{\partial^2 \psi}{\partial m_L \partial \kappa} < 0$. Therefore, higher κ's lead to a higher marginal cost of memory manipulation ($-m_L \geq 0$). Then, differentiation of the condition (1.8) and an inspection of the condition for boundary equilibria establishes the second and third claims. ∎

Proof of Proposition 6: Follows from equations (1.10) and (1.11). ∎

Proof of Proposition 7: First, consider the forgetfulness model. From Corollary 1,

$$E[u] - U(\Sigma_a) = (1 - q_a)\psi_L(m_L^*(a)) \geq 0.$$

Then, Proposition 5 implies that the amount of manipulation $|m_L^*|$ is increasing in Δu. Since, by Assumption 3, $\Delta u(\kappa, a)$ is decreasing in κ for any a, it follows that $E[u] - U(\Sigma_a)$ is decreasing in κ.

Consider the limited memory model. From Corollary 1,

$$E[u] - U(\Sigma_a) = q_a \psi_H(m_H^*(a)) \geq 0.$$

It can be shown that the set of equilibrium manipulations is increasing in the benefit of manipulation Δu (in the sense of strong set order). Then, Assumption 3 and the monotonicity of $\psi_H(m_H)$ in $m_H \geq 0$ imply that the set of equilibrium premia $\{E[u] - U(\Sigma_a)\}$ is decreasing in κ (in the sense of strong set order). ■

Proof of Proposition 8: The representation follows equations (1.13) and (1.15). Note that $q = 0$ implies that, in any PBE, $m_L^* = 0$. Thus, $w(0) = 0$. Similarly, in any PBE, $q = 1$ implies $m_H^* = 0$ and, therefore, $w(1) = 1$. ■

Proof of Proposition 9: This is a special case of Proposition 14. ■

Proof of Lemma 1: Note that in this case recollections are i.i.d. Then, in order to apply Doob's consistency theorem, we need to check that there exists a set $A \in \{\varnothing, L, H\}$ such that $\theta_1 \neq \theta_2 \implies \Pr_{\theta_1}(A) \neq \Pr_{\theta_2}(A)$. In each period, the probability of each recollection $\hat{\sigma}$ (which are i.i.d.) is

$$\begin{aligned}
\Pr(\hat{\sigma} = H|\theta) &= \Pr(\sigma = H|\theta) \times \eta_H, \\
\Pr(\hat{\sigma} = L|\theta) &= [1 - \Pr(\sigma = H|\theta)]\eta_L, \\
\Pr(\hat{\sigma} = \varnothing|\theta) &= \Pr(\sigma = H|\theta)(1 - \eta_H) + [1 - \Pr(\sigma = H|\theta)](1 - \eta_L).
\end{aligned}$$

Since $\Pr(\sigma = H|\theta)$ is strictly increasing in θ, it follows that $\theta_1 > \theta_2$ implies $\Pr_{\theta_1}(\hat{\sigma} = H) > \Pr_{\theta_2}(\hat{\sigma} = H)$ and $\Pr_{\theta_1}(\hat{\sigma} = L) < \Pr_{\theta_2}(\hat{\sigma} = L)$, which verifies the condition. ■

Proof of Lemma 2: To simplify the notation, consider the distribution of q instead of the distribution of θ. This is without loss of generality since $q = \Pr(\sigma = H|\theta)$ is strictly increasing in θ. With some abuse of notation, I will write $F(q|h^n)$ for the c.d.f. of q given history h^n.

Note that actions $b_n \in B$ are functions of the sequence of recollections $\{\hat{\sigma}_1, ..., \hat{\sigma}_n\}$. Therefore, to simplify notation and with no loss of generality, I omit the actions $\{b_1, b_2, ..., b_n\}$ from histories. Thus, with some abuse of notation, I will refer to a history as a sequence of recollections $h^n = \{\hat{\sigma}_1, ..., \hat{\sigma}_n\}$ in

all the proofs in the appendix.

Denote by $h^{n\backslash k}$ the history $\{\hat{\sigma}_1, ..., \hat{\sigma}_{k-1}, \hat{\sigma}_{k+1}, ..., \hat{\sigma}_n\}$. I will use the following result:

Claim A2. For any history h^n, we have:

$$F\left(q|h^{n\backslash k}, \hat{\sigma}_k = H\right) \leq F\left(q|h^{n\backslash k}, \hat{\sigma}_k = L\right).$$

This claim states that, for any history, a high signal is good news about q and a low signal is bad news about q in terms of first-order stochastic dominance.

Note that the p.d.f. conditional on h^n is

$$f(q|h^n) = \frac{\begin{array}{c}\prod\limits_{t:\sigma_t=\varnothing}\left[q\left(1-\eta_H-m^*_{H,t}\right)+(1-q)\left(1-\eta_L-m^*_{L,t}\right)\right] \\ \times \prod\limits_{t:\sigma_t=H} q\left(\eta_H+m^*_{H,t}\right) \times \prod\limits_{t:\sigma_t=L}(1-q)\left(\eta_L+m^*_{L,t}\right) \times f(q)\end{array}}{\int\left\{\begin{array}{c}\prod\limits_{t:\sigma_t=\varnothing}\left[q\left(1-\eta_H-m^*_{H,t}\right)+(1-q)\left(1-\eta_L-m^*_{L,t}\right)\right] \\ \times \prod\limits_{t:\sigma_t=H} q\left(\eta_H+m^*_{H,t}\right) \times \prod\limits_{t:\sigma_t=L}(1-q)\left(\eta_L+m^*_{L,t}\right) \times f(q)\end{array}\right\} dq}.$$

Let $\#H$ denote the number of times that a signal $\hat{\sigma} = H$ was recollected: $\#\{t : \hat{\sigma}_t = H\}$. Similarly, define $\#L$ as $\#\{t : \hat{\sigma}_t = L\}$.[62] Then, after some algebraic manipulations, we can write:

$$f(q|h^n) = \frac{\prod\limits_{t:\sigma_t=\varnothing}\left[q\left(1-\eta_H-m^*_{H,t}\right)+(1-q)\left(1-\eta_L-m^*_{L,t}\right)\right] \times q^{\#H} \times (1-q)^{\#L} \times f(q)}{\int \prod\limits_{t:\sigma_t=\varnothing}\left[q\left(1-\eta_H-m^*_{H,t}\right)+(1-q)\left(1-\eta_L-m^*_{L,t}\right)\right] \times q^{\#H} \times (1-q)^{\#L} \times f(q)\, dq}.$$

Note that $f(q|h^n)$ is not a function of $m_{H,t}$ and $m_{L,t}$ for any history h^n such that $\hat{\sigma}_t \neq \varnothing$. This follows from the signals σ_t being i.i.d. and the fact that $\hat{\sigma}_t = \sigma_t$ when $\hat{\sigma}_t \neq \varnothing$. Integrating the equation above, we obtain

$$F(x|h^n) = \frac{\int_0^x \prod\limits_{t:\sigma_t=\varnothing}\left[q\left(1-\eta_H-m^*_{H,t}\right)+(1-q)\left(1-\eta_L-m^*_{L,t}\right)\right] \times q^{\#H} \times (1-q)^{\#L} \times f(q)\, dq}{\int_0^1 \prod\limits_{t:\sigma_t=\varnothing}\left[q\left(1-\eta_H-m^*_{H,t}\right)+(1-q)\left(1-\eta_L-m^*_{L,t}\right)\right] \times q^{\#H} \times (1-q)^{\#L} \times f(q)\, dq}.$$

We are now ready to prove the Claim:

[62] Obviously, $\#H$ and $\#L$ are functions of histories. We omit this dependence for notational clarity.

Proof of Claim A2. We have to show that

$$\frac{\int_0^x \prod_{t.\sigma_t=\varnothing} \left[q\left(1-\eta_H-m^*_{H,t}\right)+(1-q)\left(1-\eta_L-m^*_{L,t}\right)\right] \times q^{\#H} \times (1-q)^{\#L} \times f(q)\,dq}{\int_0^1 \prod_{t:\sigma_t=\varnothing} \left[q\left(1-\eta_H-m^*_{H,t}\right)+(1-q)\left(1-\eta_L-m^*_{L,t}\right)\right] \times q^{\#H} \times (1-q)^{\#L} \times f(q)\,dq} \leq$$
$$\frac{\int_0^x \prod_{t:\sigma_t=\varnothing} \left[q\left(1-\eta_H-m^*_{H,t}\right)+(1-q)\left(1-\eta_L-m^*_{L,t}\right)\right] \times q^{\#H-1} \times (1-q)^{\#L+1} \times f(q)\,dq}{\int_0^1 \prod_{t\cdot\sigma_t=\varnothing} \left[q\left(1-\eta_H-m^*_{H,t}\right)+(1-q)\left(1-\eta_L-m^*_{L,t}\right)\right] \times q^{\#H-1} \times (1-q)^{\#L+1} \times f(q)\,dq}.$$

When $x=0$, both sides become 0 and, when $x=1$, both sides are equal to 1.

The derivative of the LHS with respect to x is

$$\frac{\prod_{t\cdot\sigma_t=\varnothing} \left[x\left(1-\eta_H-m^*_{H,t}\right)+(1-x)\left(1-\eta_L-m^*_{L,t}\right)\right] \times x^{\#H} \times (1-x)^{\#L} \times f(x)}{\int_0^1 \prod_{t:\sigma_t=\varnothing} \left[q\left(1-\eta_H-m^*_{H,t}\right)+(1-q)\left(1-\eta_L-m^*_{L,t}\right)\right] \times q^{\#H} \times (1-q)^{\#L} \times f(q)\,dq},$$

and the derivative of the RHS with respect to x is

$$\frac{\prod_{t:\sigma_t=\varnothing} \left[x\left(1-\eta_H-m^*_{H,t}\right)+(1-x)\left(1-\eta_L-m^*_{L,t}\right)\right] \times x^{\#H-1} \times (1-x)^{\#L+1} \times f(x)}{\int_0^1 \prod_{t:\sigma_t=\varnothing} \left[q\left(1-\eta_H-m^*_{H,t}\right)+(1-q)\left(1-\eta_L-m^*_{L,t}\right)\right] \times q^{\#H-1} \times (1-q)^{\#L+1} \times f(q)\,dq}.$$

Note that $\frac{dRHS}{dq} > \frac{dLHS}{dq}$ if and only if

$$\frac{\prod_{t:\sigma_t=\varnothing} \left[x\left(1-\eta_H-m^*_{H,t}\right)+(1-x)\left(1-\eta_L-m^*_{L,t}\right)\right] \times x^{\#H} \times (1-x)^{\#L} \times f(x)}{\int_0^1 \prod_{t:\sigma_t=\varnothing} \left[q\left(1-\eta_H-m^*_{H,t}\right)+(1-q)\left(1-\eta_L-m^*_{L,t}\right)\right] \times q^{\#H} \times (1-q)^{\#L} \times f(q)\,dq}$$
$$\leq \frac{\prod_{t:\sigma_t=\varnothing} \left[x\left(1-\eta_H-m^*_{H,t}\right)+(1-x)\left(1-\eta_L-m^*_{L,t}\right)\right] \times x^{\#H-1} \times (1-x)^{\#L+1} \times f(x)}{\int_0^1 \prod_{t:\sigma_t=\varnothing} \left[q\left(1-\eta_H-m^*_{H,t}\right)+(1-q)\left(1-\eta_L-m^*_{L,t}\right)\right] \times q^{\#H-1} \times (1-q)^{\#L+1} \times f(q)\,dq}.$$

Rearranging, we obtain:

$$\frac{x}{\int_0^1 \prod_{t:\sigma_t=\varnothing} \left[q\left(1-\eta_H-m^*_{H,t}\right)+(1-q)\left(1-\eta_L-m^*_{L,t}\right)\right] \times q^{\#H} \times (1-q)^{\#L} \times f(q)\,dq}$$
$$\leq \frac{(1-x)}{\int_0^1 \prod_{t:\sigma_t=\varnothing} \left[q\left(1-\eta_H-m^*_{H,t}\right)+(1-q)\left(1-\eta_L-m^*_{L,t}\right)\right] \times q^{\#H-1} \times (1-q)^{\#L+1} \times f(q)\,dq}.$$

Thus, $\frac{dRHS}{dq} > \frac{dLHS}{dq}$ if and only if

$$\rho(x) > \frac{\int_0^1 \prod_{t:\sigma_t=\varnothing} \left[q\left(1-\eta_H-m^*_{H,t}\right)+(1-q)\left(1-\eta_L-m^*_{L,t}\right)\right] \times q^{\#H} \times (1-q)^{\#L} \times f(q)\,dq}{\int_0^1 \prod_{t:\sigma_t=\varnothing} \left[q\left(1-\eta_H-m^*_{H,t}\right)+(1-q)\left(1-\eta_L-m^*_{L,t}\right)\right] \times q^{\#H-1} \times (1-q)^{\#L+1} \times f(q)\,dq},$$

where $\rho(x) = \frac{x}{1-x}$. Since $\rho(0) = 0$, $\rho(1) = +\infty$, $\rho(x)$ is strictly increasing in x, and the term on the right is a positive constant, there exists a unique $\bar{x}$ such that

$$\rho(x) > (<) \frac{\int_0^1 \prod_{t:\sigma_t=\varnothing} \left[q\left(1-\eta_H-m^*_{H,t}\right)+(1-q)\left(1-\eta_L-m^*_{L,t}\right)\right] \times q^{\#H} \times (1-q)^{\#L} \times f(q)\,dq}{\int_0^1 \prod_{t:\sigma_t=\varnothing} \left[q\left(1-\eta_H-m^*_{H,t}\right)+(1-q)\left(1-\eta_L-m^*_{L,t}\right)\right] \times q^{\#H-1} \times (1-q)^{\#L+1} \times f(q)\,dq},$$

if $x < (>)\,\bar{x}$.

Therefore, we have that $\frac{dRHS}{dq} > \frac{dLHS}{dq}$ if $x < \bar{x}$ and $\frac{dRHS}{dq} < \frac{dLHS}{dq}$ if $x > \bar{x}$. Thus, the inequality is satisfied for all q (it is satisfied with strict inequality whenever $q \in (0,1)$ and with equality at $q \in \{0,1\}$.

■

Now, we are ready to prove the lemma:

Proof of Lemma 2. As shown previously, $F(x|h^n)$ is not a function of $m^*_{L,k}$ and $m^*_{H,k}$ for k such that $\hat{\sigma}_k \neq \varnothing$. Therefore, we only need to establish the results for k such that $\hat{\sigma}_k = \varnothing$.

Consider an arbitrary k such that $\hat{\sigma}_k = \varnothing$. Then, $F(x|h^n)$ is equal to

$$\frac{\begin{array}{l}\left(1-\eta_H-m^*_{H,k}\right)\int_0^x q^{\#H+1}\times(1-q)^{\#L}\prod_{t\neq k:\sigma_t=\varnothing}\left[q\left(1-\eta_H-m^*_{H,t}\right)+(1-q)\left(1-\eta_L-m^*_{L,t}\right)\right]f(q)\,dq \\ +\left(1-\eta_L-m^*_{L,k}\right)\int_0^x q^{\#H}\times(1-q)^{\#L+1}\prod_{t\neq k:\sigma_t=\varnothing}\left[q\left(1-\eta_H-m^*_{H,t}\right)+(1-q)\left(1-\eta_L-m^*_{L,t}\right)\right]f(q)\,dq\end{array}}{\begin{array}{l}\left(1-\eta_H-m^*_{H,k}\right)\int_0^1 q^{\#H+1}(1-q)^{\#L}\prod_{t\neq k:\sigma_t=\varnothing}\left[q\left(1-\eta_H-m^*_{H,t}\right)+(1-q)\left(1-\eta_L-m^*_{L,t}\right)\right]f(q)\,dq \\ +\left(1-\eta_L-m^*_{L,k}\right)\int_0^1 q^{\#H}\times(1-q)^{\#L+1}\prod_{t\neq k:\sigma_t=\varnothing}\left[q\left(1-\eta_H-m^*_{H,t}\right)+(1-q)\left(1-\eta_L-m^*_{L,t}\right)\right]f(q)\,dq\end{array}}.$$

With some algebraic manipulations, it follows that $\frac{dF}{dm_{H,k}}(x|h^n) > 0$ if and only if

$$\frac{\int_0^x q^{\#H} \times (1-q)^{\#L+1} \prod_{t \neq k:\sigma_t = \varnothing} \left[q\left(1-\eta_H - m^*_{H,t}\right) + (1-q)\left(1-\eta_L - m^*_{L,t}\right)\right] f(q)\, dq}{\int_0^1 q^{\#H} \times (1-q)^{\#L+1} \prod_{t \neq k:\sigma_t = \varnothing} \left[q\left(1-\eta_H - m^*_{H,t}\right) + (1-q)\left(1-\eta_L - m^*_{L,t}\right)\right] f(q)\, dq} >$$
$$\frac{\int_0^x q^{\#H+1} \times (1-q)^{\#L} \prod_{t \neq k:\sigma_t = \varnothing} \left[q\left(1-\eta_H - m^*_{H,t}\right) + (1-q)\left(1-\eta_L - m^*_{L,t}\right)\right] f(q)\, dq}{\int_0^1 q^{\#H+1} \times (1-q)^{\#L} \prod_{t \neq k:\sigma_t = \varnothing} \left[q\left(1-\eta_H - m^*_{H,t}\right) + (1-q)\left(1-\eta_L - m^*_{L,t}\right)\right] f(q)\, dq}$$

Note that the left-hand side is equal to $F(x|h^n, \hat{\sigma}_{n+1} = L)$, whereas the right-hand side is equal to $F(x|h^n, \hat{\sigma}_{n+1} = H)$. From the previous claim, it follows that $F(x|h^n, \hat{\sigma}_{n+1} = L) \geq F(x|h^n, \hat{\sigma}_{n+1} = H)$, which proves that the condition above is satisfied. Therefore, we have shown that $\frac{dF}{dm^*_{H,k}}(x|h^n) > 0$. The argument for $\frac{dF}{dm^*_{L,k}}(x|h^n) < 0$ is analogous. ∎

Proof of Proposition 10: The result is immediate from inequality 1.19, Lemma 1, and the fact that the sets of histories with zero measure are the same for all relevant manipulation efforts. ∎

Proof of Proposition 11: In period N, conditions 1 and 2 from the definition of a PBE state that

$$\begin{aligned} m_{L,N}\left(L, h^{N-1}\right) \in & \arg\max_{m_L} (\eta_L + m_L) \int u(\theta)\, dF\left(\theta|L, h^{N-1}\right) \\ & + (1 - \eta_L - m_L) \int u(\theta)\, dF\left(\theta|\varnothing, h^{N-1}\right) - \psi_L(m_L), \end{aligned}$$

and

$$m_{L,N}\left(H, h^{N-1}\right) \in \arg\max_{m_H} \left\{ \begin{array}{c} (\eta_H + m_H)\left[\int u(\theta)\, dF\left(\theta|H, h^{N-1}\right)\right] \\ + (1 - \eta_H - m_H)\left[\int u(\theta)\, dF\left(\theta|\varnothing, h^{N-1}\right)\right] - \psi_H(m_H) \end{array} \right\}.$$

From Proposition 10, it follows that $\int u(\theta)\, dF\left(\theta|h^N\right)$ converges to $u(\theta)$ for almost all histories. But, when $\int u(\theta)\, dF\left(\theta|h^N\right) = u(\theta)$, it follows that $m_L\left(L, h^{N-1}\right)$ maximizes

$$(\eta_L + m_L)\, u(\theta) + (1 - \eta_L - m_L)\, u(\theta) - \psi_L(m_L) = u(\theta) - \psi_L(m_L),$$

which has a global maximum at $m_L = 0$. Hence, by continuity, it follows that $m_L\left(L, h^{N-1}\right) \to 0$ (a.s.). Similarly, when $\int u(\theta)\, dF\left(\theta|h^N\right) = u(\theta)$, $m_H\left(H, h^{n-1}\right)$ maximizes $u(\theta) - \psi_H(m_H)$ so that $m_H\left(H, h^{n-1}\right) \to 0$ (a.s.). ∎

Proof of Proposition 12: Given $\sigma = s$, self 1 maximizes

$$(\eta_s + m_s)(\theta_s + \pi_s) + (1 - \eta_s - m_s)[\alpha^* \theta_H + (1 - \alpha^*)\theta_L + \pi_s] - \psi_s(m_s),$$

where $\alpha^* = \alpha(m_L^*, m_H^*)$. Simplifying, this expression becomes:

$$(\eta_s + m_s)\theta_s + (1 - \eta_s - m_s)[\alpha^* \theta_H + (1 - \alpha^*)\theta_L] + \pi_s - \psi_s(m_s).$$

Therefore, the solution of the maximization program of self 1 is independent of π_s. It thus follows the set of manipulation efforts $m_s^*(T)$ that are part of a PBE is the same for all π_L and π_H, $s \in \{H, L\}$.

The self 1 chooses $a = T$ if

$$q\pi_H + (1 - q)\pi_L \geq q\psi_H(m_H^*(T)) + (1 - q)\psi_L(m_L^*(T)).$$

The result then follows from the fact that the left-hand side is not a function of π_H and π_L. ∎

Proof of Proposition 13: The expected utility of self 1 if she chooses (I, NE) is $q(\theta_H + \pi_H) + (1 - q)(\theta_L + \pi_L)$. Her expected utility if she chooses NI is $q\theta_H(1 - q)\theta_L$. Because $q\pi_H + (1 - q)\pi_L > 0$, it follows that NI is never chosen.

If self 1 chooses (I, E), she obtains:

$$q(\theta_H + \pi_H) + (1 - q)\theta_L - q\psi_H(m_H^*) - (1 - q)\psi_L(m_L^*).$$

Therefore, (I, E) is chosen if

$$q(\theta_H + \pi_H) + (1 - q)\theta_L - q\psi_H(m_H^*) - (1 - q)\psi_L(m_L^*) \geq q(\theta_H + \pi_H) + (1 - q)(\theta_L + \pi_L).$$

Rearranging, we obtain

$$-(1 - q)\pi_L \geq q\psi_H(m_H^*) + (1 - q)\psi_L(m_L^*). \tag{1.28}$$

Proceeding as in the proof of Proposition 12, it can be shown that the set of manipulation efforts $m_s^*(I, E)$ that are part of a PBE is the same for all π_L and π_H, $s \in \{H, L\}$. Then, the result follows immediately from equation (1.28). ∎

Proof of Proposition 14: For any PBE, define the expected manipulation cost as $MC(\varepsilon) \equiv q\psi_H(m_H^*(\varepsilon)) +$

$(1-q)\psi_H(m_L^*(\varepsilon))$. Note that $\lim_{\varepsilon\to 0_+}\chi(\varepsilon H,\varepsilon L)=0$. Therefore, for small ε, equation (1.17) becomes:

$$\int u(\theta,CE(\varepsilon))\,dF(\theta)=qu_H(\varepsilon H)+(1-q)\,u_L(\varepsilon L)-MC(\varepsilon). \tag{1.29}$$

Since $MC(0)>0$ and, by the Theorem of the Maximum, $MC(\varepsilon)$ is continuous, it follows that $MC(\varepsilon)>0$ for small ε. Hence, $\lim_{\varepsilon\to 0_+}MC(\varepsilon)>0$. Then, equation (1.29) yields:

$$\lim_{\varepsilon\to 0_+}\int u(\theta,CE(\varepsilon))\,dF(\theta)>qu_H(0)+(1-q)\,u_L(0)=\int u(\theta,0)\,dF(\theta),$$

where the last equality follows from Bayes' rule. Since u is continuous and increasing in money, this implies that $\lim_{\varepsilon\to 0_+}CE(\varepsilon)>0$. Hence, $\lim_{\varepsilon\to 0_+}\pi(\varepsilon)/\varepsilon=-\lim_{\varepsilon\to 0_+}CE(\varepsilon)/\varepsilon<0$. ∎

Proof of Proposition 15: Since $MC(\varepsilon)=0$ for all ε, equation (1.17) becomes

$$\int u(\theta,CE(0))\,dF(\theta)=qu_H(\varepsilon H)+(1-q)\,u_L(\varepsilon L)+z\chi(\varepsilon H,\varepsilon L). \tag{1.30}$$

Substituting $\chi(0,0)=0$, yields

$$\int u(\theta,CE(0))\,dF(\theta)=qu_H(0)+(1-q)\,u_L(0).$$

Therefore, Bayes' rule implies that $\int u(\theta,CE(0))\,dF(\theta)=\int u(\theta,0)\,dF(\theta)$ and, because u is strictly increasing in money, $\pi(0)=-CE(0)=0$.

Differentiating equation (1.30), it follows that

$$CE'(0)=\frac{qu'_H(0)\,H+(1-q)\,u'_L(0)\,L+z(H-L)\left[u'_H(0)-u'_L(0)\right]}{qu'_H(CE)+(1-q)\,u'_L(CE)}.$$

Substituting $qH+(1-q)L=0$, yields

$$CE'(0)=K\left[u'_H(0)-u'_L(0)\right],$$

where $K=\frac{H}{qu'_H(0)+(1-q)u'_L(0)}\left(q+\frac{z}{1-q}\right)>0$. Thus, applying L'Hospital, we obtain

$$\lim_{\varepsilon\to 0_+}\pi(\varepsilon)/\varepsilon=-CE'(0)=-K\left[u'_H(0)-u'_L(0)\right],$$

which concludes the proof. ∎

D1 Proofs of Remarks and Examples

Proof of the claim in Remark 1: Let $\hat{\mu}$ and μ denote the cumulative distribution functions of $\hat{\theta}_{\hat{\sigma}} \in \left\{\hat{\theta}_L, \hat{\theta}_{\varnothing}, \hat{\theta}_H\right\}$ and $\theta_\sigma \in \{\theta_L, \theta_H\}$, respectively. $\hat{\theta}_{\hat{\sigma}}$ second-order stochastically dominates θ_σ if, for any concave function $g : \Theta \to \mathbb{R}$,

$$\int g\left(\hat{\theta}_{\hat{\sigma}}\right) d\mu\left(\hat{\theta}_{\hat{\sigma}}\right) \geq \int g\left(\theta_\sigma\right) d\mu\left(\theta_\sigma\right). \tag{1.31}$$

But

$$\int g\left(\theta_\sigma\right) d\mu\left(\theta_\sigma\right) = qg\left(\theta_H\right) + (1-q)\, g\left(\theta_L\right), \text{ and}$$

$$\begin{aligned}\int g\left(\hat{\theta}_{\hat{\sigma}}\right) d\mu\left(\hat{\theta}_{\hat{\sigma}}\right) &= q\left(m_H + \eta_H\right) g\left(\theta_H\right) + \left[q\left(1 - m_H - \eta_H\right) + (1-q)\left(1 - m_L - \eta_L\right)\right] g\left(\hat{\theta}_{\varnothing}\right) \\ &\quad + (1-q)\left(\eta_L + m_L\right) g\left(\theta_L\right).\end{aligned}$$

Substituting in inequality (1.31) and dividing by $q\left(1 - m_H - \eta_H\right) + (1-q)\left(1 - m_L - \eta_L\right)$, we obtain:

$$g\left(\alpha\left(m_L, m_H\right)\theta_H + \left[1 - \alpha\left(m_L, m_H\right)\right]\theta_L\right) \geq \alpha\left(m_L, m_H\right) g\left(\theta_H\right) + \left[1 - \alpha\left(m_L, m_H\right)\right] g\left(\theta_L\right),$$

which is true because g is concave. ∎

Example 5: It is helpful to separate the analysis in 2 cases: (i) $q \geq \frac{2}{5}$, and (ii) $q < \frac{2}{5}$. In case (i), self 2 chooses a high ex-post action, $b(\varnothing) = b_H$ when she expects self 1 to manipulate her memory, $m_L = -\frac{2}{3}$. In case (ii), she chooses a low ex-post action, $b(\varnothing) = b_L$ when she expects $m_L = -\frac{2}{3}$.

Case (i):

The DM chooses to manipulate her memory if

$$\begin{aligned}&\left(1 - \frac{2}{3}\right)\left[v_L\left(b_L\right) + \tau\left(L\right)\right] \\ &\qquad + \frac{2}{3}\left\{\alpha\left[v_H\left(b_L\right) + \tau\left(L\right)\right] + (1-\alpha)\left[v_L\left(b_L\right) + \tau\left(L\right)\right]\right\} \\ &\qquad\qquad - \psi_L\left(-\frac{2}{3}\right) > v_L\left(b_L\right) + \tau\left(L\right),\end{aligned}$$

where α denotes the weight implies by Bayes' rule. This inequality is satisfied if and only if $\alpha > \frac{3}{32}$. Substituting the definition of α, we obtain $q > \frac{2}{31}$,which is satisfied since $q \geq \frac{2}{5} > \frac{2}{31}$.

The ex-ante expected utility from the signal is thus

$$q\left[v_H\left(b_H\right)+\tau\left(H\right)\right]+\frac{2}{3}\left(1-q\right)\left[v_L\left(b_H\right)+\tau\left(L\right)\right]$$
$$+\left(1-q\right)\left(1-\frac{2}{3}\right)\left[v_L\left(b_L\right)+\tau\left(L\right)\right]-\left(1-q\right)\psi_L\left(-\frac{2}{3}\right),$$

which is equal to $\frac{83q+1}{12}$. If the DM makes an uninformed decision, she obtains an ex-ante utility of

$$\begin{aligned} q\left[v_H\left(b_H\right)+\tau\left(H\right)\right]+\left(1-q\right)\left[v_L\left(b_H\right)+\tau\left(L\right)\right] \text{ if } b &= b_H, \text{ and} \\ q\left[v_H\left(b_L\right)+\tau\left(H\right)\right]+\left(1-q\right)\left[v_L\left(b_L\right)+\tau\left(L\right)\right] \text{ if } b &= b_L. \end{aligned}$$

Thus, her utility is $7q$ if $q \geq \frac{1}{2}$, and $5q+1$ if $q < \frac{1}{2}$. The surplus from observing the signal is then

$$\frac{83q+1}{12}-\max\left\{7q,5q+1\right\}=\begin{cases} \frac{1-q}{12} \text{ if } q \geq \frac{1}{2}, \\ \frac{23q-11}{12} \text{ if } q < \frac{1}{2} \end{cases},$$

which is positive if and only if $q > \frac{11}{23}$.

Case (ii): In this case, because $b\left(\varnothing\right)=b_L$, the signal has no value. Therefore, the DM is always (weakly) better off by not observing the signal. In particular, since she exerts memory manipulation if $q \geq \frac{2}{31}$, the surplus is strictly negative for $q > \frac{2}{31}$ and it is equal to zero if $q < \frac{2}{31}$.

Example 6: Following the same steps as in Example 5, it is straightforward to show that the surplus from observing the signal S is equal to

$$\begin{aligned} \left(1-q\right)\left(\frac{3}{4}-\bar{m}\right) \text{ if } q &\geq \frac{1}{2}, \text{ and} \\ 2q-1+\left(1-q\right)\left(1-\bar{m}-\frac{1}{4}\right) \text{ if } q &< \frac{1}{2}. \end{aligned}$$

Therefore, S is always positive when $q \geq \frac{1}{2}$. Furthermore, for $q < \frac{1}{2}$, it is positive if and only if

$$2q-1+\left(1-q\right)\left(1-\bar{m}-\frac{1}{4}\right) \geq 0,$$

which simplifies to

$$\frac{3}{4}-\frac{1-2q}{1-q} \geq \bar{m}.$$

Noting that $\bar{m}$ is distributed according to the c.d.f. Φ concludes the proof.

References

Akerlof, G. A. and W. T. Dickens (1982). "The Economic Consequences of Cognitive Dissonance," American Economic Review, 72(3), 307-319.

Allport, G. W. (1943). "The ego in contemporary psychology," Psychological Review, 50, 451-478.

Anderson, J. R. (1995). *Learning and Memory: An Integrated Approach,* New York: John Wiley & Sons.

Arkes, H. R., and C. Blumer (1985). "The psychology of sunk cost," Organizational Behavior and Human Decision Processes, 35, 124-140.

Baron, R. A. (1999). "Counterfactual thinking and venture formation: The potential effects of thinking about what might have been," Journal of Business Venturing 15, 79–91.

Bell, D. E. (1982). "Regret in decision making under uncertainty," Operations Research, 30, 961-981.

Benabou, R. (2008). "Groupthink: Collective Delusions in Organizations and Markets," Mimeo., Princeton University.

Benabou, R. and J. Tirole (2002). "Self-Confidence and Personal Motivation," Quarterly Journal of Economics, 117(3), 871-915.

Benabou, R. and J. Tirole (2004). "Willpower and Personal Rules," Journal of Political Economy, 112 (4), 848-886.

Benabou, R. and J. Tirole (2006a). "Belief in a Just World and Redistributive Politics," Quarterly Journal of Economics, 121(2), 699-746.

Benabou, R. and J. Tirole (2006b). "Incentives and Prosocial Behavior," American Economic Review, 96(5), 1652-1678.

Benabou, R. and J. Tirole (2006c). "Identity, Dignity and Taboos: Beliefs as Assets," Mimeo., Princeton University and Université de Toulouse.

Berglas, S. and Baumeister, R. (1993). *Your Own Worst Enemy: Understanding the Paradox of Self-Defeating Behavior*, New York: BasicBooks.

Bernheim, B. D. and R. Thomadsen (2005). "Memory and Anticipation," Economic Journal, 115, 271–304.

Billot, A, Chateauneuf, A., Gilboa, I., and Tallon, J.-M. (2000). "Sharing Beliefs: Between Agreeing and Disagreeing," Econometrica, 68(3), 685-694.

Bodner, R. and D. Prelec. "Self-signaling in a neo-Calvinist model of everyday decision making," in Psychology and Economics, Vol. II. Brocas and J. Carillo (eds.), Oxford University Press, 2002.

Brocas, I. and J. D. Carrillo (2008). "The Brain as a Hierarchical Organization," American Economic Review, forthcoming.

Brockner, J. (1992). "The escalation of commitment to a failing course of action: toward theoretical

progress," Academy of Management Review, 17, 39–61.

Brockner, J., Houser, R., Birnbaum, G., Lloyd, K., Deitcher, J., Nathanson, S., and J. Z. Rubin (1986). "Escalation of Commitment to an Ineffective Course of Action: The Effect of Feedback Having Negative Implications for Self-Identity," Administrative Science Quarterly, 31, 109-126.

Brunnermeier, M. K., and J. A. Parker (2005). "Optimal Expectations," American Economic Review, 95, 1092-1118.

Budescu, D. V., Weinberg, S., and T. S. Wallsten (1988). "Decisions based on numerically and verbally expressed uncertainties," Journal of Experimental Psychology: Human Perception and Performance, 14, 281-294.

Budescu, D. V., Kuhn, K. M., Kramer, K. M., and T. R. Johnson (2002). "Modeling certainty equivalents for imprecise gambles," Organizational Behavior and Human Decision Processes, 88, 748-768.

Byrne, C. C. and Kurland, J. A. (2001). "Self-deception in an Evolutionary Game," Journal of Theoretical Biology, 212 (4), 457-480.

Camerer, C. F. and R. A. Weber (1999). "The econometrics and behavioral economics of escalation of commitment: a re-examination of Staw and Hoang's NBA data," Journal of Economic Behavior & Organization, 39, 59-82.

Caplin, A. and K. Eliaz (2003). "AIDS Policy and Psychology: A Mechanism-Design Approach," RAND Journal of Economics, 34(4), 631-646.

Caplin, A. and J. Leahy (2001). "Psychological Expected Utility Theory and Anticipatory Feelings," Quarterly Journal of Economics, 116 (1), 55-79.

Caplin, A. and J. Leahy (2004). "The Supply of Information by a Concerned Expert," Economic Journal, 114, 487-505.

Compte, O., and and Andrew Postlewaite. (2004) "Confidence-Enhanced Performance," American Economic Review, 94, 1536-1557.

Chau, A. W., & Phillips, J. G. (1995) "Effects of perceived control upon wagering and attributions in computer blackjack," Journal of General Psychology, 122, 253–269.

Chow, C. C. and R.K. Sarin (2001). "Comparative Ignorance and the Ellsberg Paradox," Journal of Risk and Uncertainty, 22, 129-139.

Dawson, E., Savitsky, K., and Dunning, D. (2006). "Don't Tell Me, I Don't Want To Know: Understanding People's Reluctance to Obtain Medical Diagnostic Information," Journal of Applied Social Psychology, 36, 751-768.

Di Mauro, C. (2008). "Uncertainty Aversion vs. Competence: An Experimental Market Study," Theory and Decision, 64, 301-331.

Dow, J. (1991). "Search Decisions with Limited Memory," Review of Economic Studies, 58, 1-14.

Dunning, D. (1995). "Trait importance and modifiability as factors influencing self-assessment and self-enhancement motives," Personality and Social Psychology Bulletin, 21, 1297-1306.

Eliaz, K. and R. Spiegler (2006). "Can anticipatory feelings explain anomalous choices of information sources?" Games and Economic Behavior 56, 87–104.

Epstein, L. G. (1999). "A Definition of Uncertainty Aversion," Review of Economic Studies, 66, 579-608.

Epstein, L. G. (2007). "Living with Risk," Rochester Center for Economic Research, Working Paper #534.

Falk, A., Huffman, D., and U. Sunde (2006). "Self-Confidence and Search," IZA Discussion Paper #2525.

Festinger, L. (1957). *A theory of cognitive dissonance.* Stanford, CA: Stanford University Press.

Fox, C. R., and A. Tversky, "Ambiguity Aversion and Comparative Ignorance," Quarterly Journal of Economics, 110, 585-603.

Fox, C. R., and M. Weber. (2002). "Ambiguity aversion, comparative ignorance, and decision context," Organizational Behavior and Human Decision Processes, 88, 476-498.

Fudenberg , D. and D. K. Levine (2006). "A Dual Self Model of Impulse Control," American Economic Review, 96, 1449-1476.

Fudenberg , D. and D. K. Levine (2007). "Self Control, Risk Aversion, and the Allais Paradox," Working Paper, Harvard University and Washington University in St. Louis.

Genesove, D. and C. Mayer (2001). "Loss Aversion And Seller Behavior: Evidence From The Housing Market," Quarterly Journal of Economics, 116, 1233-1260.

Gollwitzer, P. M., Earle, W. B., and Stephan, W. G. (1982). "Affect as a determinant of egotism: Residual excitation and performance attributions," Journal of Personality and Social Psychology, 43, 702-709.

Goodie, A. S. (2003). "The Effects of Control on Betting: Paradoxical Betting on Items of High Confidence With Low Value," Journal of Experimental Psychology: Learning, Memory, and Cognition, 29, 598–610.

Goodie, A. S. and D. L. Young (2007). "The skill element in decision making under uncertainty: Control or competence?," Judgment and Decision Making, 2, 189–203.

Greenwald, A. G. (1980). "The Totalitarian Ego: Fabrication and Revision of Personal History," American Psychologist 35, 603-618.

Grieco, D. and R. M. Hogarth (2004). "Excess entry, ambiguity seeking, and competence: An experimental investigation," Working Papers #778, Universitat Pompeu Fabra.

Hanany, E. and P. Klibanoff (2007). "Updating preferences with multiple priors," Theoretical Economics, 2, 261-298.

Harbaugh, R. (2008). "Prospect Theory or Skill Signaling?" Working Paper, Indiana University.

Heath, C. and A. Tversky (1991). "Preference and belief: Ambiguity and competence in choice under uncertainty," Journal of Risk and Uncertainty, 4(1), 5-28.

Hellman, M. E. and T. M. Cover (1973). "A Review of Recent Results on Learning with Finite Memory," Problems of Control and Information Theory, 221-227.

Hilgard, E. R. (1949). "Human motives and the concept of self. American Psychologist," 4, 374-382.

Hirshleifer, D. and I. Welch (2002). "An Economic Approach to the Psychology of Change: Amnesia, Inertia, and Impulsiveness," Journal of Economics & Management Strategy, 11, 379 - 421.

Horswill, M. S. and F. P. McKenna (1999). "The effect of perceived control on risk taking," Journal of Applied Social Psychology, 29, 378-392.

Hvide, H. (2002). "Pragmatic beliefs and overconfidence," Journal of Economic Behavior and Organization, 2002, 48, 15-28.

Johnston, W. A. (1967). "Individual performance and self-evaluation in a simulated team," Organizational Behavior and Human Performance, 2, 309-328.

Josephs, R. A., Larrick, R. P., Steele, C. M., and Nisbett, R. E. (1992). "Protecting the Self from the Negative Consequences of Risky Decisions," Journal of Personality and Social Psychology, 62, 26-37.

Kahneman, D. and A. Tversky, 1979. "Prospect theory: An analysis of decision under risk," Econometrica, 47, 263–291.

Kahneman, D., J. L. Knetsch., and R. H. Thaler (1990). "Experimental Tests of the Endowment Effect and the Coase Theorem," Journal of Political Economy, 98(6), 1325-1348.

Keppe, H. and M. Weber (1995). "Judged Knowledge and Ambiguity Aversion," Theory and Decision, 39, 51-77.

Kihlstrom, J. F., Beer, J. S., and S. B. Klein (2003). "Self and identity as memory," In: M. R. Leary and J. P. Tangney (Eds.), Handbook of Self and Identity. (Guilford Press: New York, NY).

Kilka, M. and M. Weber (2000). "Home Bias in International Stock Return Expectation," Journal of Psychology and Financial Markets 1, 176-192.

Korner, I. (1950) "Experimental Investigation of Some Aspects of the Problem of Repression: Repressive Forgetting." New York, NY: Contributions to Education, No. 970, Bureau of Publications, Teachers' College, Columbia University.

Kőszegi, B. (2003) "Health anxiety and patient behavior," Journal of Health Economics, 22, 1073–1084.

Kőszegi, B. (2006) "Ego Utility, Overconfidence, and Task Choice," Journal of the European Economic Association, 4, 673-707.

Kuehberger, A. and Perner, J. (2003). "The role of competition and knowledge in the Ellsberg task," Journal of Behavioural Decision Making, 16, 181–191.

Kunda, Z. and Sanitioso, R. (1989) "Motivated Changes in the Self–Concept," Journal of Personality and Social Psychology, 61, 884–897.

Larrick, R. P. (1993) "Motivational Factors in Decision Theories: The Role of Self-Protection," Psychological Bulletin, 113, 440-450.

Li, W. (2007). "Changing One's Mind when the Facts Change: Incentives of Experts and the Design of Reporting Protocols," Review of Economic Studies, 74, 1175–1194.

List, J. (2003). "Does Market Experience Eliminate Market Anomalies?," Quarterly Journal of Economics, 118 (1), pages 41-71.

List, J. and Haigh, M. S. (2005). "A simple test of expected utility theory using professional traders," Proceedings of the National Academy of Sciences 102, 945-948.

Loomes, G. and Sugden, R. (1982). "Regret theory: An alternative theory of rational choice under uncertainty," Economic Journal, 92, 805-824.

Lowenstein, G. A. (1987), "Anticipation and the valuation of delayed consumption," Economic Journal, 97, 666-684.

Mather, M., E. Shafir, and M.K. Johnson (2003). "Memory and Remembering chosen and assigned options," Memory & Cognition, 31 (3), 422-433.

Mill J. (1829). *Analysis of the Phenomena of the Human Mind*, (London: Baldwin and Craddock).

Mukerji, S. (1998). "Ambiguity Aversion and Incompleteness of Contractual Form," American Economic Review, 88 (5),1207-1231.

Mullainathan, S. (2002). "A memory-based model of bounded rationality," Quarterly Journal of Economics, 117(3), 735-774.

Philipson, T. and R. Posner (1995). "A Theoretical and Empirical Investigation of the Effects of Public Health Subsidies for STD Testing," Quarterly Journal of Economics, 110(2), 445-474.

Piccione, M. and Rubinstein, A. (1997). "On the interpretation of decision problems with imperfect recall," Games and Economic Behavior, 20, 3–24.

Prelec, D. (1998). "The probability weighting function," Econometrica, 66, 497–527.

Prelec, D. (2008). "Self-delusion: A neuroeconomics model and fMRI evidence," Seminar on the Foundations of Human Social Behavior, University of Zurich.

Quattrone G. A., and Tversky, A. (1984). "Causal Versus Diagnostic Contingencies: On Self-Deception and the Voter's Illusion." Journal of Personality and Social Psychology, 46, 237-248.

Rabin, M. (1994). "Cognitive Dissonance and Social Change," Journal of Economic Behavior and Organization, 23, 177-194.

Rabin, M. (2000). "Risk Aversion and Expected-Utility Theory: A Calibration Exercise," Econometrica, 68(5), 1281-1292.

Rapaport, D. (1961). *Emotions and Memory*, (New York: Science Editions).

Robbins, H. (1956). "A Sequential Decision Problem with a Finite Memory," Proceedings of the National Academy of Sciences, 42, 920-923.

Safra, Z. and Segal, U. (2008). "Calibration Results for Non-Expected Utility Theories," Econometrica, forthcoming.

Samuelson, P. A. (1963). "Risk and uncertainty: A fallacy of large numbers," Scientia, 6th Series 57, 1-6.

Segal, U. and A. Spivak (1990). "First Order versus Second Order Risk Aversion," Journal of Economic Theory, 51, 111-125.

Schelling, T. (1985). "The Mind as a Consuming Organ," In: J. Elster, ed. The Multiple Self. New York: Cambridge University Press.

Sedikides, C., Green, J. D., and Pinter, B. T. (2004). "Self-protective memory," In: D. Beike, J. Lampinen, and D. Behrend, eds., The self and memory. Philadelphia: Psychology Press.

Staw, B. M. (1976) "Knee Deep in the Big Muddy," Organizational Behavior and Human Decision Process 35, 124-140.

Staw, B. M. (1981). "The Escalation of Commitment to a Course of Action," The Academy of Management Review, 6, 577-587.

Steele, C. M. (1988). "The Psychology of Self-Affirmation: Sustaining the Integrity of the Self," In: L. Berkowitz (Ed.), *Advances in experimental social psychology,* (Vol. 21, 261-302). New York: Academic Press.

Taylor, K. (1995) "Testing Credit and Blame Attributions as Explanation for Choices under Ambiguity," Organizational Behavior and Human Decision Processes, 64, 128-137.

Taylor, S. E. and Gollwitzer, P. M. (1995). "The effects of mindset on positive illusions," Journal of Personality and Social Psychology, 69, 213-226.

Thaler, R.H. (1980). "Toward a Positive Theory of Consumer Choice," Journal of Economic Behaviour and Organization, 1, 39-60.

Thaler, R.H., and H.M. Shefrin (1981). "An Economic Theory of Self-control," Journal of Political Economy, 89, 392-406.

Tirole, J. (2008). "Cognition and Incomplete Contracts," American Economic Review, forthcoming.

Trivers, R, (2000). "The Elements of a Scientific Theory of Self-Deception," Annals of the New York Academy of Sciences, 907, 114-131.

Tversky, A. and D. Kahneman (1992). "Advances in prospect theory: cumulative representation of uncertainty," Journal of Risk and Uncertainty, 5, 297–323.

Van den Steen, E. (2004). "Rational Overoptimism (and Other Biases)," American Economic Review,

94, 1141-1151.

Weinberg, B. A. (2006). "A Model of Overconfidence." Mimeo., Ohio State University.

Wilson, A. (2003). "Bounded Memory and Biases in Information Processing," Mimeo., Princeton University.

Zuckerman, M. (1979). "Attibution of success and failure revisited, or: The motivational bias is alive and well in attribution theory," Journal of Personality, 47, 245-287.

Chapter 2

Competition over time-inconsistent consumers

2.1 Introduction

A significant amount of evidence suggests that, in some markets, consumers are not time-consistent. When making intertemporal decisions, they tend to give higher relative weights to an earlier future period as it gets closer. Consumers also often overestimate their degree of time-consistency.

DellaVigna and Malmendier (2006), for example, analyze health clubs that offer both monthly contracts and 10-visit passes. Most consumers take the monthly contracts even though they would spend less if they bought a 10-visit pass. Shui and Ausubel (2005) consider a randomized experiment in the credit card industry. Consumers are offered a contract with a lower interest rate and a shorter duration and one with a higher interest rate but a longer duration. Although most consumers take the contract with a shorter duration, they would pay less if they chose the longer duration one because they continue to borrow on the credit card.[1]

How do firms respond to time-inconsistency among consumers? DellaVigna and Malmendier

[0]This chapter was previously publised on the *Journal of Public Economic Theory*. See Gottlieb (2008).

[1]Oster and Scott-Morton (2005) consider the relation between newsstand and subscription prices of magazines. They show that magazines associated with future benefits ("investment magazines") have relatively higher subscription costs than those with immediate benefits ("leisure magazines"), indicating that firms may be exploiting the consumers' time-inconsistency.

(2004) have shown that firms set prices above marginal cost for goods with immediate rewards and deferred costs (*leisure goods*) and below marginal cost for goods with immediate costs and deferred rewards (*investment goods*) both in the monopolistic and in the competitive case. This result follows from the fact that commitment devices are valued by consumers and, therefore, firms are able to benefit from providing them. Furthermore, as long as consumers are aware of being time-inconsistent, the equilibrium is efficient.

Although DellaVigna and Malmendier (2004) show that their model is consistent with evidence from investment goods such as the health club industry and the vacation time-sharing industry as well as evidence from leisure goods such as the credit card and the mobile phone industries, it is not consistent with observations from the tobacco, alcohol, and unhealthy food industries (which are probably among the most compelling examples of markets with time-inconsistent consumers). Since these are leisure goods, the model implies that consumers would receive a lump-sum transfer and pay a marginal price higher than the marginal cost. However, we do not observe such lump-sum transfers to consumers in these three examples.

In this essay, I argue that an important component missing from the DellaVigna and Malmendier model is the fact that firms are typically unable to offer exclusive contracts. In some markets (such as the cigarettes, alcohol, and unhealthy foods markets), the cost for the firm of enrolling a potential user or for the consumer to switch between different firms is usually very low and, therefore, contracts are non-exclusive. Then, if a firm offers a positive lump-sum transfer to consumers and charges high marginal prices (as happens in the case of leisure goods with exclusivity), another firm may benefit from selling the good at a lower price after the contract has been signed. Hence, the impossibility to prevent a consumer from buying from another firm after the contract has been signed is an important restriction for leisure goods.

In equilibrium, there is a stark asymmetry between leisure and investment goods. While the equilibrium in the investment goods case is the same as in DellaVigna and Malmendier (2004) and there is no role for taxation when consumers are aware of their time-inconsistency, in the leisure goods case prices are equal to marginal cost and the equilibrium is always inefficient. The key idea is that time-inconsistent consumers face an ex-post incentive to circumvent the prearranged commitment devices. In the case of leisure goods, firms are able to profit ex-post

by offering these contracts, which breaks down the equilibrium.[2] In the case of investment goods, they are not able to profit and the equilibrium with commitment devices is maintained. This prediction seems to be consistent with evidence that markets for leisure goods with small costs of enrolling new users and small switching costs do not feature lump-sum transferences to consumers.[3]

An implication of the model is that, in the case of leisure goods, increased competition may decrease total surplus. If a monopolist could commit not to renegotiate the contract but firms could not prevent consumers from signing contracts with other firms, total surplus in leisure goods markets would be higher under monopoly than under competition.[4]

The efficient allocation in the case of leisure goods can be obtained by a sales tax that corrects for the "internalities" of consuming the good. It does not depend on the degree of naiveté and requires a relatively small amount of information from a regulator. On the other hand, the optimal tax in the case of investment goods depends on the consumers' naiveté (the tax is zero if they are sophisticated) and requires much more information. Therefore, my model suggests that there is a much larger role for government intervention in leisure goods markets than in investment goods markets.

This paper is related to a growing literature on behavioral industrial organization.[5] Gabaix and Laibson (2006) analyze information shrouding by firms serving both rational and myopic consumers. They show that, under some conditions, informational shrouding persists even in competitive environments. Spiegler (2006) assumes that consumers use a sampling procedure because they have a limited ability to evaluate complex objects. Then, increased competition leads to more obfuscation instead of the convergence of prices to marginal costs.[6]

---

[2] In that sense, my model is similar to the models of commitment as renegotiation-proofness of Hart and Tirole (1988) and Dewatripont (1988). However, the issue here is not the incentive to renegotiate the contract but to accept another contract from a different firm.

[3] In an independent work, Köszegi (2005) has recently provided a similar suggestion for why we do not observe lump-sum transfers in some markets for leisure goods He does not, however, explicitly model competition. He also does not study the welfare properties of the resulting equilibrium. Nevertheless, as in the model presented in this paper, he also obtains an asymmetry between markets for leisure and investment goods.

[4] This result relies on the assumption that consumers' preferences are homogeneous and, therefore, there is no deadweight loss from monopoly. If consumers were heterogeneous, the result would depend on whether the deadweight loss from monopoly would overweight the gain from providing commitment.

[5] Ellison (2006) provides a survey of this literature.

[6] Eliaz and Spiegler (2006) consider a model where a monopolist faces consumers with limited ability to predict changes in their future tastes. Heidhues and Köszegi (2005) study the optimal contract offered by a monopolist facing loss-averse consumers.

The chapter is also related to the literature on the optimal regulation of goods consumed by time-inconsistent agents. Gruber and Köszegi (2001) extended the Becker and Murphy's (1988) rational-addiction model to the case of quasi-hyperbolic consumers. O'Donoghue and Rabin (2003) studied the optimal taxation of unhealthy goods based on Ramsey's commodity taxation model. The main conclusion of these papers is that it is optimal to tax leisure goods and to subsidize investment goods.[7] The main reason for taxing leisure goods is the provision of "internalities": Taxes provide incentives for consumers to act according to their long-run preferences. In other words, taxing immediate-rewards goods provides a commitment device that avoids consumers from felling tempted to behave differently from how their long-run selves would act.

In both papers, however, it is assumed that prices are set equal to the marginal cost in the absence of regulation. However, as shown by DellaVigna and Malmendier (2004), in the presence of time-inconsistent consumers, firms do not set prices equal to marginal cost. In this essay, I consider the optimal taxes when prices are endogenously set by competitive firms when contracts are nonexclusive.

2.2 The model

The model is a competitive version of DellaVigna and Malmendier (2004). There are three periods. In the first period, consumers make a take-it-or-leave-it offer of a contract to firms. Because the good is indivisible, there is no loss of generality in assuming that a first-period contract consists of a two-part tariff (L, p), where L denotes a lump-sum price and p is a usage price. Both L and p are paid in period 2.

Consumption occurs in period 2. The good provides an immediate payoff (in period 2) of $-c$ and a delayed payoff (in period 3) of b. If $-c < 0 < b$, the good generates an initial cost and a delayed benefit. These goods are called *investment goods.* If $b < 0 < -c$, then the good generates an immediate benefit and a delayed cost. These are called *leisure goods.*

Consumers and firms have a common prior about F, the distribution of c. F is assumed to

[7] Gruber and Köszegi have estimated the optimal cigarete taxes based on their model. They found that the provision of "internalities" due to hyperbolic discounting leads to an optimal tax of at least $1 per pack more than the traditional model.

be a twice continuously differentiable distribution function with strictly positive density f on $\Theta \subset \mathbb{R}$. We take $\Theta = \mathbb{R}_+$ for the case of investment goods ($-c < 0 < b$) and $\Theta = \mathbb{R}_-$ for the case of leisure goods ($b < 0 < -c$).

In period 2, the realization of the immediate cost c is observed. Consumers make a take-it-or-leave-it offer of a contract to firms. A second-period contract is a price $\hat{p}$ contingent on consuming the good.[8] Then, they decide whether to consume the good.

If consumption occurred in period 2, consumers get a payoff with expected value b in period 3. Since there are no choices in this period, it is irrelevant whether b is deterministic or stochastic.[9]

Therefore, the timing of the game is as follows:

1. Consumers offer a two-part tariff (L, p) to firms, where L denotes a lump-sum price and p is a usage price (both paid in period 2).

2. The immediate cost c is drawn from the distribution F. Consumers offer a price $\hat{p}$ contingent on consuming the good. Then, they decide whether to consume the good.

3. If the good was consumed in period 2, the consumer gets a payoff with expected value b.

Consumers have quasi-hyperbolic preferences:

$$U_t = u_t + \beta \sum_{s>t} \delta^{s-t} u_s.$$

They are *exponential* (or time-consistent) when their time-consistency parameter β is equal to 1. If β is less than 1, they are *hyperbolic* (or time-inconsistent). A *partially naive* hyperbolic agent has true time-consistency parameter β, but believes that in the future she will behave like a hyperbolic agent with parameter $\hat{\beta} \in [\beta, 1]$. When $\hat{\beta} = \beta$, the agent is *sophisticated.*

Each firm faces a cost of providing the good equal to $a > 0$, incurred in period 2 (which is when production occurs).[10] We assume that the firm has access to a credit market and faces a

[8] There is no loss of generality on assuming that the price if the good is not consumed is zero.

[9] Notice that delayed benefits b do not depend on the realization of c.

[10] We assume that there are no costs of signing a contract. Therefore, the firm faces no cost if a consumer signs a contract but does not consume the good. Our results still hold if these costs are small. On the other hand, assuming the presence of costs of signing a contract is equivalent to ruling out non-exclusive contracts if the costs are high enough.

discount factor of δ.

Note that the offer of the two-part tariff is made under symmetric information. Therefore, a time-consistent consumer would choose marginal cost pricing and would extract all profits through the fixed fee (which, in this case, leads to $L = 0, p = a$). The equilibrium in the general case is obtained by backward induction.

If a consumer has not accepted a contract in the first period, her second period program (conditional on consuming the good) consists on choosing the price $\hat{p}$ that maximizes her utility subject to leaving non-negative profits to the firms:

$$\max_{\hat{p}} \beta\delta b - \hat{p} - c$$
$$s.t. \qquad \hat{p} \geq a.$$

The unique solution to this program is $\hat{p} = a$. If she has accepted a first period contract (L, p), her program (conditional on consuming the good) is

$$\max_{\hat{p}} \beta\delta b - \min\{\hat{p}, p\} - c$$
$$s.t. \qquad \hat{p} \geq a.$$

The solution of this program is $\hat{p} = a$ if $p > a$ and $\hat{p} \geq a$ otherwise, and the consumer prefers the second period contract if $p > a$.

If a consumer accepts a first-period contract and consumes the good in the second period, she expects to obtain a benefit of $\hat{\beta}\delta b$ and faces a cost of $c + p$ in the second period. Hence, she expects to consume the good with probability $F\left(\hat{\beta}\delta b - p\right)$ and gets an expected utility of $\beta\delta\left[\pi(p) - L\right]$, where $\pi(p) := \int_{-\infty}^{\hat{\beta}\delta b - p} (\delta b - p - c)\, dF(c)$ is the net expected value of consumption.

The fact that the consumer prefers the second period contract if $p > a$ places an important restriction on the set of contracts that can be demanded in the first period. Since the she will not consume using a contract (L, p) with $p > a$, the firm gets negative profits when $L < 0$. Therefore, we cannot have contracts with $p > a$ and $L < 0$.

Clearly, a consumer would never offer a first period contract with $L > 0$ and $p > a$ since it would leave strictly positive profits to the firm (she could improve by offering $L' = 0$ and the

same usage price). Hence, the possibility of offering contracts in the second period implies that first-period contracts must satisfy $p \leq a$.

Conversely, the consumer prefers to use a first-period contract whenever $p < a$. Hence, when $p < a$, the zero-profit condition for the firm becomes

$$L + F(\beta\delta b - p)(p - a) \geq 0.$$

Thus, any contract such that $p \leq a$ and $L + F(\beta\delta b - p)(p - a) \geq 0$ is accepted by the firms.

Hence, conditional on accepting a first period contract, the consumer's first period program is[11]

$$\begin{aligned} &\max_{p,L} \beta\delta [\pi(p) - L] \\ &s.t. \qquad p \leq a \\ &L + F(\beta\delta b - p)(p - a) \geq 0. \end{aligned}$$

For the moment, ignore the $p \leq a$ constraint. Then, the first-order condition yields:

$$p - a = -\delta b \left(1 - \hat{\beta}\right) \frac{f(\hat{\beta}\delta b - p)}{f(\beta\delta b - p)} - \frac{F(\hat{\beta}\delta b - p) - F(\beta\delta b - p)}{f(\beta\delta b - p)}, \tag{2.1}$$

which is the same expression as in DellaVigna and Malmendier (2004). The per-unit price p differs from the marginal cost a for two reasons. First, time-inconsistent consumers use prices as commitment devices as long as they are not fully naive (i.e. $\hat{\beta} < 1$). This is reflected by the term $-\delta b \left(1 - \hat{\beta}\right) \frac{f(\hat{\beta}\delta b - p)}{f(\beta\delta b - p)}$. Second, by underestimating the probability of consuming the good, the consumer does not take into account the full cost of higher marginal tariffs (captured by the second term).

In the case of investment goods, the term on the right of equation (2.1) is negative. Therefore, the $p \leq a$ constraint does not bind and the equilibrium is exactly the same as the one in

[11] I assume that aggregate surplus is strictly concave for $p \leq a$. A sufficient condition is that $\frac{f'(c)}{f(c)}(\delta b - a - c) < 1$ for all $c \geq \beta\delta b - a$. Since the consumer can choose $L = 0$, $p = a$, there is no loss of generality in assuming that the first period contract is accepted.

DellaVigna and Malmendier. The lump-sum price is then given by the zero-profit condition:

$$\begin{aligned} L &= -F(\beta\delta b - p)(p - a) \\ &= F(\beta\delta b - p)\left[\delta b\left(1 - \hat{\beta}\right)\frac{f(\hat{\beta}\delta b - p)}{f(\beta\delta b - p)} + \frac{F(\hat{\beta}\delta b - p) - F(\beta\delta b - p)}{f(\beta\delta b - p)}\right]. \end{aligned}$$

In the case of leisure goods, the term is positive. Hence, the constraint binds and the solution is $p = a$. The zero-profit condition implies that the lump-sum price is $L = 0$. Therefore, *the equilibrium in the case of leisure goods involves marginal cost pricing.*

Notice that, because leisure goods are priced at marginal cost, the model implies that prices are uninformative about the consumers' degree of time-inconsistency when contracts are nonexclusive.

The model appears to be compatible with evidence from several markets. As DellaVigna and Malmendier (2004) argue, there is evidence of below marginal cost pricing in investment goods markets such as health clubs, vacation time-sharing. On the other hand, leisure goods such as tobacco, alcohol, and unhealthy food do not feature ex-ante lump-sum transfers to consumers and higher unit prices.

An important assumption of the model is that consumers are able to buy from other firms in the period that consumption occurs. If contracts could only be signed in a period before consumption occurs, nonexclusivity would not lead to marginal cost pricing. In the credit card industry, for example, there is a time lag between obtaining a credit card and being able to use it. Therefore, even though credit cards are clearly nonexclusive, usage prices are higher than marginal costs.

2.3 Welfare Analysis

This Section characterizes the optimal taxes for leisure and investment goods. I will follow most of the literature in treating the agent's long-run time preferences as the relevant for social welfare.[12] Moreover, since partially naive consumers have a mistaken perception about their

[12]This view is defended by O'Donoghue and Rabin (1999) and Bernheim and Rangel (2005). An alternative approach is to apply a Pareto criterion [see, e.g., Laibson, Repetto, and Tobacman (1998) or Diamond and Köszegi (2003)]. A problem with the Pareto optimality approach is that it leads to an incomplete ranking.

true time-consistency parameter, I will attribute the correct parameter β for the preferences used in welfare comparisons.

Proposition 18 *Suppose consumers are sophisticated. Then the equilibrium is Pareto efficient in the case of investment goods ($b > 0$) and Pareto inefficient in the case of leisure goods ($b < 0$).*

Proof. The first part follows from the fact that, in the case of investment goods, the equilibrium maximizes the consumer's (long-run) utility subject to firms not getting negative profits. The second part follows from the fact that, by the strict concavity of total surplus, the unique maximum is given by equation (2). Since profits are equal to zero and the consumer is worse off under marginal-cost pricing, it follows that the equilibrium is Pareto inefficient. ■

The inefficiency in the case of leisure goods stems from the fact that, with marginal-cost pricing, consumers do not internalize the full effects of consuming the good. Since the consumers internalize only a fraction β of future costs $-\delta b$, they do not account for the remaining part: $-\delta b(1-\beta)$. Therefore, an optimal tax should increase the price of the good by $-\delta b(1-\beta)$ so that the perceived costs equal perceived benefits. Since this tax raises the consumers' long-run utility and does not change the firms' profit (which remains equal to zero), the resulting allocation Pareto dominates the equilibrium without taxes.

For investment goods, the equilibrium is Pareto inefficient when consumers are partially naive because they do not take into account the true parameter β.[13] In order to obtain the optimal sales tax in the case of investment goods, define the after-tax marginal cost as $\hat{a} \equiv a+\tau$. Then, the optimal tax should be such that after-tax prices are equal to the optimal price $a - \delta b(1-\beta)$. Hence,

$$p = \hat{a} - \delta b\left(1-\hat{\beta}\right)\frac{f(\hat{\beta}\delta b - p)}{f(\beta\delta b - p)} - \frac{F(\hat{\beta}\delta b - p) - F(\beta\delta b - p)}{f(\beta\delta b - p)} = a - \delta b(1-\beta)$$

Expressing in terms of τ, we obtain

$$\tau = \delta b\left[\left(1-\hat{\beta}\right)\frac{f(\hat{\beta}\delta b - p)}{f(\beta\delta b - p)} - (1-\beta)\right] - \frac{F(\hat{\beta}\delta b - p) - F(\beta\delta b - p)}{f(\beta\delta b - p)}.$$

[13] More precisely, for generic distributions F, the equilibrium in the case of investment goods is Pareto inefficient.

I formally state the results above in the following proposition:

Proposition 19 *Suppose consumers are partially naive. Then, the equilibrium is Pareto inefficient. A Pareto efficient allocation can be obtained by imposing a sales tax of:*

(i) $\tau = -\delta b\,(1-\beta) > 0$ *in the case of case of leisure goods; and*

(ii) $\tau = \delta b\left[\left(1-\hat{\beta}\right)\frac{f(\hat{\beta}\delta b-p)}{f(\beta\delta b-p)} - (1-\beta)\right] - \frac{F(\hat{\beta}\delta b-p)-F(\beta\delta b-p)}{f(\beta\delta b-p)}$ *in the case of investment goods.*

The allocation implemented by the sales tax maximizes the consumer surplus subject to the firms obtaining nonnegative profits.

Notice that the optimal sales tax for leisure goods is always positive. It is a function of the long-run discount factor δ, the time-inconsistency parameter β, and the expected delayed costs b of consuming the good and does not depend on the distribution of the immediate benefits c from consuming.[14]

The optimal sales tax for investment goods may be either positive or negative. Furthermore, it requires not only knowledge of the long-run discount factor δ, the expected delayed costs b of consuming the good, and the time-inconsistency parameter β, but also the distribution of c and the perceived parameter of time-inconsistency $\hat{\beta}$. Hence, determining the optimal tax for these goods requires knowledge of several parameters which are hard to be estimated.

2.4 Conclusion

Although the DellaVigna and Malmendier model appears to be successful in explaining evidence for the health club, vacation time-sharing, and credit cards industries, it is not compatible with evidence from some of the most compelling examples of markets featuring time-inconsistent consumers: tobacco, alcohol, and unhealthy food. According to their model, firms should be offering ex-ante lump-sum transfers to consumers and higher unit prices so that consumers purchase exclusively from them.

[14] Gruber and Köszegi (2001) provide estimates of all the parameters required in order to calculate the optimal tax in our model for the case of cigarettes.

I have shown that the evidence can be explained if one drops the exclusivity assumption, which seems to be unreasonable in these markets. While in the exclusive-contracts case the difference between leisure and investment goods is only a sign change in prices, there is an important asymmetry between them when contracts are non-exclusive.

The marginal-cost-pricing result for the case of leisure goods suggests that taxes should focus on leisure goods such as cigarettes, alcohol, and unhealthy food and not on investment goods, where the market may provide commitment devices efficiently (at least if consumers are sophisticated; otherwise, the optimal tax may be positive or negative and depends on parameters that are hard to estimate).

References

BECKER, G. S. and K. M. MURPHY (1988) A Theory of Rational Addiction, *Journal of Political Economy* **96**, 675–700.

BERNHEIM, B. D. and A. RANGEL (2005) Behavioral Public Economics: Welfare and Public Policy Analysis with Non-Standard Decision Makers, *NBER Working Paper number 11518*

DELLAVIGNA, S. and U. MALMENDIER (2004) Contract Design and Self-Control: theory and evidence, *Quarterly Journal of Economics* **119**, 353–402.

DELLAVIGNA, S. and U. MALMENDIER (2006) Paying Not To Go To The Gym, *American Economic Review* **96**, 694-719.

DEWATRIPONT, M. (1988) Commitment Through Renegotiation-Proof Contracts with Third Parties, *Review of Economic Studies* **55**, 377-90.

DIAMOND, P. and B. KÖSZEGI (2003) Quasi-Hyperbolic Discounting and Retirement, *Journal of Public Economics* **87**, 1839-1872.

ELIAZ, K. and SPIEGLER, R. (2006) Contracting with Diversely Naïve Agents, *Review of Economic Studies* **73**, 689-714.

ELLISON, G. (2006) Bounded Rationality in Industrial Organization, Blundell, *Advances in Economics and Econometrics: Theory and Applications*, by R. Persson and W. K. Newey, Eds., Cambridge University Press: Cambridge, 143-174.

GABAIX, X. and D. LAIBSON (2006) Shrouded Attributes, Consumer Myopia, and Information Suppression in Competitive Markets, *Quarterly Journal of Economics* **121**, 505–40.

GOTTLIEB, D. (2008) Competition Over Time-Inconsistent Consumers, *Journal of Public Economic Theory* **10**, 673-684.

GRUBER, J. and B. KÖSZEGI (2001) Is Addiction 'Rational'? Theory and Evidence, *Quarterly Journal of Economics* **116**, 1261–1303.

HART, O. and J. TIROLE (1988) Contract Renegotiation and Coasian Dynamics, *Review of Economic Studies* **55**, 509-40.

HEIDHUES, P. and B. KÖSZEGI (2005) The Impact of Consumer Loss Aversion on Pricing, *University of Bonn and Berkeley University.*

KÖSZEGI, B. (2005) On the Feasibility of Market Solutions to Self-Control Problems, *Swedish Economic Policy Review* **12**, 71-94.

LAIBSON, D. I. (1997) Golden Eggs and Hyperbolic Discounting, *Quarterly Journal of Economics* **112**, 443–477.

LAIBSON, D. I., A. REPETTO, and J. TOBACMAN (1998) Self-Control and Saving for Retirement, *Brookings Papers on Economic Activity* **1,** 91-196.

O'DONOGHUE, T. D., and M. RABIN (1999) Doing It Now or Later, *American Economic Review* **89**, 103–124.

O'DONOGHUE, T. D., and M. RABIN (2003) Studying Optimal Paternalism, Illustrated by a Model of Sin Taxes, *American Economic Review* **93**, 186-191.

OSTER, S. M. and F. M. SCOTT MORTON (2005) Behavioral Biases Meet the Market: The Case of Magazine Subscription Prices, Advances in Economic Analysis & Policy **5**.

PHELPS, E. S. and R. A. POLLAK (1968) On Second-Best National Saving and Game-Equilibrium Growth, *Review of Economic Studies* **35**, 185-199.

RAMSEY, F. P. (1927) A Contribution to the Theory of Taxation, *Economic Journal* **37**, 47-61.

SHUI, H. and L. M. AUSUBEL (2005) Time Inconsistency in the Credit Card Market, *University of Maryland.*

SPIEGLER, R. (2006) Competition over agents with boundedly rational expectations, *Theoretical Economics* **1**, 207–231.

Chapter 3

A Model of Mixed Signals with Applications to Countersignaling

3.1 Introduction

In the initial papers in the signaling literature, the informational asymmetry consists of a unidimensional parameter, which is known to only one side of the market [e.g. Spence, 1973]. Then, under the natural condition that individuals can be ordered according to their marginal utility of signaling (single-crossing property), there exists a family of separating equilibria in which signals reveal information monotonically. In the job market models, for example, higher education discloses information about higher productivity. These equilibria are ranked by the Pareto optimality criterion; moreover, only the Pareto dominant equilibrium is robust to competition among firms [Riley, 1979].

More recently, scholars have identified conditions under which the main results from unidimensional models extend to multi-dimensional ones. These conditions typically involve some form of separability between dimensions. However, in many situations, a single instrument conveys information about multiple characteristics. In such cases, good and bad characteristics may be revealed by the same instrument. We refer to a multidimensional model with a single

[0]This chapter is based on joint work with Aloisio Araujo and Humberto Moreira. It was previously published on the *RAND Journal of Economics*. See Araujo, Gottlieb, and Moreira (2007).

signaling instrument as a "mixed signals" model.[1] Another strand of literature has focused on unidimensional models featuring non-monotonic signals. These models generate "countersignaling", wherein individuals with high types choose to engage in a lower amount of signaling than medium-type individuals [Feltovich, Harbaugh, and To, 2002].

In this article, we present a two-dimensional characteristics signaling model satisfying the single-crossing property (SCP) in each dimension. Workers' characteristics are represented by a vector of cognitive and non-cognitive ability parameters. Firms can observe a combination of these characteristics through an interview but cannot precisely determine if the result of this interview was due to high cognitive or non-cognitive ability. The results of the interview process can be used to reduce the two-dimensional model to a one-dimensional model where the SCP fails to hold. Workers are able to signal their characteristics through the number of years of education they acquire. A theoretical contribution of the essay is to provide a characterization of the equilibrium in a signaling model where the SCP fails, thereby extending Araujo and Moreira's [2001] analysis of a screening model.

It is shown that countersignaling occurs whenever the schooling technology differs from the technology of firms. The model has a very intuitive testable implication: the amount of countersignaling is strictly increasing in the difference between the schooling technology and the firms' technology. Hence, countersignaling is expected to be more important in occupations that require a different combination of skills from those required in the schooling process.

This model is also employed in order to understand evidence on the General Educational Development (GED) exam. The signaling equilibrium has some interesting properties consistent with available empirical evidence on the GED: individuals with different abilities obtain the same amount of education and passing the exam signals higher cognitive skills but does not increase one's earnings. These results follow from the fact that the GED is a mixed signal: if a worker with low overall ability has passed the exam, it means that her non-cognitive ability is low. Hence, as both types of ability are used in the production process, passing the exam is not necessarily a signal of high productivity. The model suggests that the ineffectiveness of the GED exam stems from its focus on cognitive ability. A test that places a stronger emphasis on non-cognitive ability would be a more effective signal. Moreover, a simple change in the passing

[1] We have borrowed the term "mixed signal" from Cavallo, Heckman, and Hsee [1998].

standards of the GED would not affect its neutrality on wages.

3.1.1 Related Literature on Countersignaling and Mixed Signals

There are several documented examples of what appears to be countersignaling behavior. Hvide [2003] argues that intermediate individuals appear to pursue more education than bright individuals for professions where individuals without a licence are not denied work. Unlike standard signaling models of advertising predict, Clements [2004] documents that many high-quality products are sold in low-quality packages. Orzach et al. [2002] argue that, even controlling for market size, luxury cars (such as Rolls Royce and Ferrari) seem to be advertised very modestly compared to nonluxury cars.[2] In the context of fashion as a signal of status, Pinker [1999] claims that "trend-setters are members of upper classes who adopt the styles of lower classes to differentiate themselves from middle classes". According to O'Neil [2002], countersignaling led intermediately advanced countries to spend more on their military than most advanced countries after World War II.

Feltovich, Harbaugh, and To [2002] present a countersignaling model applied to the labor market. Firms access some measure of the worker's ability (which is interpreted as the recommendation of a former boss). This signal consists of the sum of the unidimensional ability of the worker and a noise term. Workers may also engage in schooling activity. If the exogenous signal were sufficiently informative about the individual productivity of workers, then it would not be profitable for them to use schooling to signal productivity. On the other hand, if the exogenous signal were completely uninformative about the workers' productivities, we would have a standard signaling model where higher types signal more. Their model can be seen as an intermediate case where the exogenous signal is sufficiently informative to separate high from medium types but not sufficient to separate medium from low types.

Our model differs from that of Feltovich, Harbaugh, and To in that uncertainty about productivity comes from the divergence between the schooling technology and the firms' technology instead of a noise term. The misalignment between these two technologies generates an incentive for some higher-productivity workers to obtain less education.

[2] Caves and Greene [1996] find no significant systematic positive correlation between quality and the amount of advertising.

Orzach, Overgaard, and Tauman [2002] present a model where firms signal product quality through prices and advertising expenditures. Product quality is represented by a parameter that may take two values. Their main conclusion is that modest advertising can be used as a signal of high quality. However, as their model features only two types of firms, they are unable to consider the emergence of non-monotone signals.

One example of mixed signals is the GED exam, which is taken by high school dropouts to certify that they have equivalent knowledge to high school graduates. The GED reveals, at the same time, high cognitive skills and low non-cognitive skills [Cameron and Heckman, 1993; Cavallo, Heckman and Hsee, 1998; and Heckman and Rubinstein, 2001]. Moreover, wages received by high school dropouts are not influenced by this certificate. Another example is the so-called "Ph.D. curse".[3] This curse refers to the difficulty of some recent Ph.D. graduates to find jobs outside of academia because firms perceive them as being too theoretically oriented and lacking enough practical abilities.[4] A third example of mixed signals is presented by Drazen and Hubrich [2003], who argue that higher interest rates show that the government is committed to maintaining a fixed exchange rate, but also signal weak fundamentals. Benabou and Tirole [2006] argue that donations signal altruism, but may also signal a desire to be perceived as altruistic.

In the labor market model, for example, an assumption of unidimensional information asymmetry implies that all relevant characteristics of an employee can be captured by a single ability-type, usually thought of as cognitive ability. However, significant empirical evidence supports the importance of non-cognitive skills as well as cognitive skills in the labor market [Heckman, Stixrud, and Urzua, 2005]. Generalization of the original results to the multidimensional case turns out not to be straightforward; in Kholleppel's [1983] example of a two-dimensional extension of Spence's model, no separating equilibrium exists.

Quinzii and Rochet [1985] and Engers [1987] provided sufficient conditions for the existence of a separating equilibrium in the multidimensional model. In Quinzii and Rochet, ability was represented by a k-dimensional vector and they assumed the existence of k (non-exclusive) different types of education. Moreover, they assumed that the signaling costs were linear and

[3] We thank Mathias Dewatripont for this example.

[4] See "Academic Careers: A Comparative Perspective," Jeroen Huisman and Jeroen Bartelse (eds), *http://www.awt.nl/uploads/files/academic.pdf.*

separable in the signals (up to a change of variables). Hence, it was as if each school required only one type of ability. Then, an individual would be able to attend a school whose system required only a type of skill (cognitive skills, for example) and another school that required only another type of skill (non-cognitive skills). Under this separability assumption (which implies that the single-crossing property holds in each dimension), Quinzii and Rochet obtain results similar to the unidimensional model: separating equilibria exist and wages are monotonic in the worker's types.

Needless to say, the educational systems assumed by Quinzii and Rochet are not realistic since all known educational systems require both cognitive and non-cognitive abilities (although in different proportions). Engers relaxed this assumption through a generalization of the unidimensional assumption that individuals' marginal utility of signaling could be ordered (single-crossing property). However, in the multidimensional case, this assumption is much less compelling since, as the number of signals rise, it becomes more likely that the single-crossing property (SCP) will not hold when one controls for one signal (i.e., the introduction of other signals may break the SCP in the multidimensional case).

The rest of the chapter is organized as follows. The basic framework is presented in Section 3.2. Section 3.3 characterizes the equilibrium. Section 3.4 studies how countersignaling may emerge and Section 3.5 employs this framework to analyze the GED exam. Section 3.6 briefly discusses examples of other environments where the model can be applied. Then, Section 3.7 concludes.

3.2 The Basic Framework

The economy consists of a continuum of informed workers who sell their labor to uninformed firms. Each worker is characterized by a two-dimensional vector of characteristics (ι, η), where ι and η represent cognitive and non-cognitive ability, respectively. For concreteness, we will refer to ι as intelligence and η as perseverance while bearing in mind that non-cognitive skills embody several other characteristics such as motivation, self-control, and other personality traits. The set of all possible characteristics is the compact set $\Theta \equiv [\iota_0, \iota_1] \times [\eta_0, \eta_1] \subset \mathbb{R}^2_{++}$ and the types are distributed according to a continuous density $p : \Theta \to \mathbb{R}_{++}$, which is assumed to be a C^2

function.

Workers are able to engage in a schooling activity $y \in \mathbb{R}_+$ which firms can observe. By engaging in such an activity, the type-(ι, η) worker incurs a cost $c(\iota, \eta, y)$. This worker's productivity depends on the vector of innate characteristics, which is not (directly) observable.

Firms have identical technologies with constant returns to scale $f(\iota, \eta)$ and act competitively.[5] Moreover, other than schooling, firms observe the result of an interview $g(\iota, \eta)$ which does not fully reveal the worker's productivity. Thus, even though firms have some idea of the overall ability of a worker, they are unable to unambiguously determine her productivity by observing the result of the interview.[6] In a more general model, we could imagine that individuals might exert effort in order to distort the market's assessment of their productivity [e.g., Holmstrom, 1999 and Dewatripont, Jewitt, and Tirole, 1999]. We have studied this possibility in a previous version of the paper, where we assumed that schooling influences the worker's performance in the interview. Most of the results presented here are unaffected.[7]

After observing schooling y and the result of the interview g, each firm offers a wage $w(y, g)$. Thus, each worker will choose the amount of schooling y in order to maximize $w(y, g) - c(\iota, \eta, y)$.

The timing of the signaling game is as follows. First, nature determines each worker's type according to the density function p. Then, workers choose their educational level contingent on their type. Subsequently, firms offer a wage $w(y, g)$ conditional on observing (y, g).

Since firms are homogeneous, we will study symmetric equilibria where the offered wage schedule is the same for every firm. As usual, we adopt the perfect Bayesian equilibrium concept:

Definition 4 *A perfect Bayesian equilibrium (PBE) for the signaling game is a profile of strategies* $\{y(\iota, g), w(y, g)\}$ *and beliefs* $\mu(\cdot \mid y, g)$ *such that*

[5] In this paper, we consider only the pure signaling case. In a previous version of the paper, we have shown that all the results also hold when schooling affects productivity [see Araujo, Gottlieb, and Moreira, 2007].

[6] The hypothesis that firms can access an additional signal that consists of a measure of the worker's ability is also present at Feltovich, Harbaugh, and To [2002].

[7] See Araujo, Gottlieb and Moreira [2007]. In this case, it can be shown that, locally, the ability to distort the result of the interview raises the amount of education in equilibrium for all individuals as in standard 'signal-jamming' models

1. *The worker's strategy is optimal given the equilibrium wage schedule:*

$$y(\iota,\eta) \in \arg\max_{\tilde{y}} w(\tilde{y}, g(\iota,\eta)) - c(\iota,\eta,\tilde{y}),$$

2. *Firms earn zero profits:* $w(y(\iota,\eta), g(\iota,\eta)) = E[f(\iota,\eta) \mid g, y]$.

3. *Beliefs are consistent:* $\mu(\iota,\eta \mid y, g)$ *is derived from the worker's strategy using Bayes' rule where possible.*

Next, we will specify the analytical forms of the functions presented.[8] The signaling technology is characterized by the following cost of signaling function:

$$c(\iota,\eta,y) = \frac{y}{\iota\eta}. \tag{3.1}$$

The function above implies that the cost of education is decreasing in intelligence and perseverance. Moreover, intelligence and perseverance are imperfect substitutes in the schooling process.[9]

We assume that the interview function is given by

$$g(\iota,\eta) = \alpha\iota + \eta, \tag{3.2}$$

where $\alpha > 0$ is the rate of substitution between perseverance and intelligence. Notice that, conditional on the interview g, the workers' types lie on the hyperplane represented by equation (3.2). Hence, by conditioning on g, the type space becomes one-dimensional. Thus, we will refer to a type-(ι,η) worker as type-$(\iota;g)$, since ι captures all private information after taking g into account.

Substituting (3.2) into (3.1), we are able to rewrite the cost of signaling as a function of the intelligence and the interview result:

$$c(\iota,g,y) = \frac{y}{\iota(g-\alpha\iota)},$$

[8] The robustness of the model to the functional forms is studied in the Appendix A.

[9] They are "cost substitutes" in the sense that $c(\iota,\eta,y)$ has increasing differences.

where we denote this function by c with some abuse of notation.

In general, the single-crossing property (SCP) may not be satisfied since

$$c_{y\iota}(\iota, g, y) = -\frac{g - 2\alpha\iota}{[\iota(g - \alpha\iota)]^2} \left\{ \begin{array}{c} > \\ < \end{array} \right\} 0 \Leftrightarrow \iota \left\{ \begin{array}{c} > \\ < \end{array} \right\} \frac{g}{2\alpha}.$$

The SCP states that the marginal utility of effort is monotonic in the worker's type. In this specific case, it means that, conditional on the interview g, more intelligence would either always decrease or always increase the cost of schooling. In particular, even for individuals with very low intelligence and very high perseverance levels, raising a unit of intelligence and decreasing α units of perseverance would decrease the marginal cost of schooling (or the opposite when the sign of $c_{y\iota}$ is reversed). Hence, the SCP is equivalent to assuming that the range of abilities is such that intelligence is always better than perseverance for schooling (or vice-versa).

The intelligence level $\iota = \frac{g}{2\alpha}$ divides the parameter space in two intervals (CS_+ and CS_-) according to the sign of $c_{y\iota}$ (negative and positive, respectively). For workers with intelligence below (above) $\frac{g}{2\alpha}$, intelligence decreases (increases) the cost of signaling given the interview result g. When the SCP is satisfied, $[\iota_0, \iota_1]$ is contained in one of these intervals. Figure 3-1 depicts a situation where $[\iota_0, \iota_1]$ overlaps these intervals.

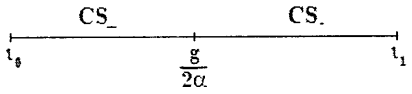

Figure 3-1: CS_+ and CS_- regions

Therefore, the two-dimensional model reduces to a one-dimensional model where the SCP may fail to hold after conditioning on g.

We assume that the worker's productivity is given by the Cobb-Douglas function

$$f(\iota, \eta) = \iota^b \eta^{1-b},$$

where $b \in (0, 1)$. If $b > \frac{1}{2}$ we say that the firm's technology is intensive in cognitive skills. Otherwise, we say that it is intensive in non-cognitive skills.

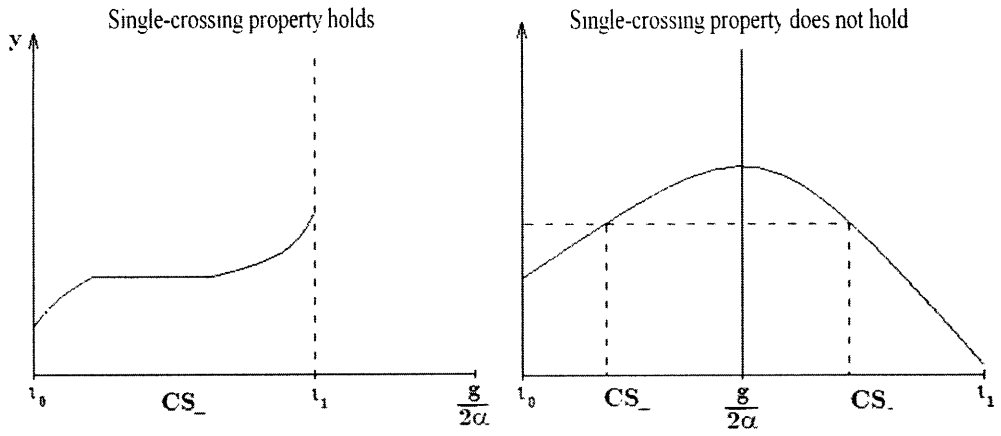

Figure 3-2: Continuous versus Discrete Pooling

It is useful to rewrite the production function conditional on the interview g as

$$s(\iota, g) = \iota^b (g - \alpha\iota)^{1-b}. \tag{3.3}$$

3.3 The Signaling Equilibria

In this section, the signaling equilibrium is characterized. First, we divide the interval of parameters in three different sets according to the degree of separation. Necessary conditions for an equilibrium are presented for each set separately. Then, we present the refinement criterion that will be employed in order to select a unique equilibrium. It consists of a generalization of Riley's [1979] criterion. Subsequently, sufficient conditions for an equilibrium are obtained.

Given an equilibrium profile of education $y(\iota, g)$, we refer to the set of types whose signal is (y, g) as the pooling set $\Theta(y, g)$. A type $(\iota; g)$ is separated if, in equilibrium, her characteristics are revealed from her signals $y(\iota, g)$ and g. If more than one type choose the same amount of education, we say that they are pooled.

In signaling models where the SCP is satisfied, incentive compatibility requires education to be increasing if $c_{y\iota} < 0$ for all ι (CS$_+$) and decreasing if $c_{y\iota} > 0$ for all ι (CS$_-$). Then, if two workers are pooled, monotonicity implies that all intermediate types must also pool with them (see graph on the left in Figure 3-2). We call these types *continuously pooled.*

When the single crossing property does not hold, incentive compatibility requires y to be increasing in the CS_+ interval and decreasing in the CS_- interval so that the equilibrium may feature non-monotone signaling. As a result, a disconnected set of workers may acquire the same level of education (see graph on the right in Figure 3-2). We say that these workers are *discretely pooled.*[10]

The precise definitions are stated below:

Definition 5 *Given an equilibrium profile of education* $\{y(\iota, g) : \iota \in [\iota_0, \iota_1],\ g \in [\alpha\iota + \eta_0,\ \alpha\iota + \eta_1]\}$*:*

1. *A type–* $(\iota; g)$ *worker is separated if* $\Theta(y(\iota, g), g) = \{(\iota, g)\}$*. A separating set is a set of types where every element is separated.*

2. *A type–* $(\iota; g)$ *worker is continuously pooled if* $\Theta(y(\iota, g), g)$ *is not discrete. A continuous pooling set is a set of types where every element is continuously pooled.*

3. *A type–* $(\iota; g)$ *worker is discretely pooled if* $\Theta(y(\iota, g), g) \neq \{(\iota, g)\}$ *is discrete. A discrete pooling set is a set of types where every element is discretely pooled.*

In any signaling equilibrium, each type must belong to one of these three sets. The selection criterion (to be discussed later) will rule out continuous pooling. Hence, the equilibrium will take a form similar to the profile on the right of Figure 3-2: There will be an interval of separated types and an interval of discretely pooled types.

In the following subsections, we study the properties of separating sets, continuously pooling sets and discrete pooling sets, respectively. As is standard in the signaling literature, the equilibrium will be represented by differential equations that follow from the worker's first-order condition. Hence, we characterize piecewise C^2 equilibria.[11]

[10] Analyzing a model of competition in many markets, Green and Laffont [1990] have also obtained an equilibrium where discrete pooling may occur. When the incumbent is able to commit to her decisions, the incentive-compatibility constraints are relaxed so that the equilibrium may feature discrete pooling. In contrast, in our paper discrete pooling occurs under the standard incentive-compatibility constraints.

[11] A piecewise C^k function is a function whose domain can be partitioned in a finite number of intervals such that the function is k times continuously differentiable in each interval. Therefore, we allow for a finite number of jumps and kinks. Our results can be generalized to piecewise C^1 functions. However, as most of the literature, we focus on the piecewise C^2 case in order to simplify the proofs.

3.3.1 Separating set

When a worker belongs to a separating set, Bayes' rule implies that belief $\mu(\iota \mid y, g)$ must be a singleton measure concentrated at ι. Then, the zero-profits condition (second condition of Definition 4) is

$$w(y(\iota, g), g) = f(\iota, g - \alpha\iota). \tag{3.4}$$

The worker's incentive-compatibility constraint is

$$\iota \in \arg\max_{\tilde{\iota}} f(\tilde{\iota}, g - \alpha\tilde{\iota}) - c(\iota, g, y(\tilde{\iota}, g)). \tag{3.5}$$

Notice that each realization of $g(\iota, \eta) = x$ defines a set of possible characteristics

$$g^{-1}(x) \equiv \{(\iota, \eta) \in [\iota_0, \iota_1] \times [\eta_0, \eta_1] : x = \alpha\iota + \eta\}.$$

As the worker's production function is a strictly concave, continuous function of ι, there exists a unique intelligence level such that the worker's productivity is maximal given the interview result g. This educational level is defined as

$$(\iota^*(g),\ \eta^*(g)) = \arg\max_{\iota,\eta} \iota^b \eta^{1-b} \quad \text{s.t. } g = \alpha\iota + \eta. \tag{3.6}$$

It follows from the first-order (necessary and sufficient) conditions of the problem above that $\iota^*(g) = \frac{bg}{\alpha}$. Hence, productivity is increasing for $\iota \leq \iota^*(g)$ and decreasing for $\iota \geq \iota^*(g)$. The interpretation of this result is straightforward. Given the result of the interview g, firms prefer moderate intelligence levels since a worker whose intelligence is too high must have a low level of perseverance.

But, as a worker must be earning her expected productivity in any separating set, it follows that wages are non-monotone in intelligence (controlling for the interview g). As shown in the previous signaling literature, when the SCP is satisfied, education is increasing in the worker's type. Suppose this is also the case when the SCP is not valid (i.e., suppose that education is increasing in intelligence). Then, firms would offer a higher salary for individuals with

intermediate schooling (as those are the most productive workers).[12] But such an allocation cannot be an equilibrium since the workers' strategies are not optimal: if they reduce the amount of schooling, their wages rise (and, of course, they obtain a higher utility). Hence, a necessary condition for incentive-compatibility is that education must be increasing in ι until $\iota^*(g)$ and decreasing after $\iota^*(g)$.

Notice that this necessary condition for an interior solution follows from the equalization between the marginal benefit from deviating and its marginal cost. Since the marginal benefit consists of the wage differential s_ι and the marginal cost consists of the marginal cost of signaling times the signaling differential $c_y y_\iota$, by computing s_ι and c_y, we obtain

$$y_\iota(\iota, g) = s(\iota, g)(bg - \alpha\iota), \tag{3.7}$$

which implies that y must be increasing if $\iota < \iota^*(g)$ and decreasing if $\iota > \iota^*(g)$.

From the local second-order condition, we obtain the usual necessary condition that education must be increasing in the CS_+ region and decreasing in CS_-. Hence, from the first- and second-order conditions of the program in equation (3.5) we obtain the following lemma, whose proof is presented in the Appendix:

Lemma 3 *In any separating set, if an education and wage profile is incentive-compatible it must satisfy*

$$y_\iota(\iota, g)(g - 2\alpha\iota) \geq 0 \tag{3.8}$$

and equation (3.7). Moreover, in a separating set, the workers with the highest level of schooling are the most productive (not the brightest or the most perseverant) and the level of schooling is strictly increasing in productivity.

Remark 9 *Notice that equation (3.8) implies that*

$$y_\iota(\iota, g) \geq 0 \Longleftrightarrow \iota \leq \frac{g}{2\alpha}. \tag{3.9}$$

Generally, equations (3.8) and (3.9) cannot hold for all ι except when $b = \frac{1}{2}$. In this case,

[12] More precisely, the wage schedule would be increasing in schooling until $y(\iota^*(g), g)$ and decreasing from that point on.

the firms' technology is identical to the signaling technology. If so, we can treat $\iota\eta$ as a single parameter and we obtain Spence's [1973] model. Moreover, education must be monotone in this (redefined) parameter.

Remark 10 *When $b \neq \frac{1}{2}$, there exists some misalignment between the firm and the worker since the relative intensity of intelligence in the schooling technology is different from that in the firm's technology. Then, if either $\frac{bg}{\alpha} \in [\iota_0, \iota_1]$ or $\frac{g}{2\alpha} \in [\iota_0, \iota_1]$, there must exist some pooling in equilibrium (since the separating set conditions cannot hold for all the interval of parameters).*

3.3.2 Continuous pooling set

Let $p(\iota \mid g)$ denote the density function of ι conditional on the result of the interview g and suppose there exists a non-degenerate closed set I which is a continuous pooling set such that no closed set $X \supsetneq I$ is a continuous pooling set. Then, $y(\iota, g) = \bar{y}(g)$ for all $\iota \in I$.

The zero-profit condition is

$$w(\bar{y}(g), g) = W(I, g), \tag{3.10}$$

where $W(I, g) \equiv \int_I f(x, g - \alpha x) p(x \mid g) dx$ is the expected productivity of a type-ι worker. Conditions 2 and 3 from Definition 4 are trivially satisfied in that given set.

3.3.3 Discrete pooling set

A distinct feature of models where the SCP does not hold is the emergence of discrete pooling, where individuals with non-adjacent types receive the same contract. This feature is a direct consequence of the possibility of non-monotone signals.

As was shown by Araujo and Moreira [2001], a necessary condition for incentive-compatibility in a discrete pooling set is the so-called marginal utility identity, according to which, if two individuals are (discretely) pooling in a contract, they should have the same marginal utility. We formally state that result as a lemma:

Lemma 4 *Suppose that $\{y(\iota, g) : \iota \in [\iota_0, \iota_1],\ g \in [\alpha\iota + \eta_0,\ \alpha\iota + \eta_1]\}$ is an incentive-compatible profile of education. If two regular workers with the same interview result choose the same level*

of education, then their marginal cost of education must be the same:

$$\left.\begin{array}{c} y(\iota, g) = y(\tilde{\iota}, g) \\ y_{\iota}(\iota, g) \neq 0 \\ y_{\iota}(\tilde{\iota}, g) \neq 0 \end{array}\right\} \Rightarrow \frac{\partial c(\iota, g, y)}{\partial y} = \frac{\partial c(\tilde{\iota}, g, y)}{\partial y}.$$

The economic interpretation of Lemma 4 is that if two non-adjacent workers with different marginal costs of education choose the same contract, one of them could benefit from deviating by choosing a different amount of schooling.

From the equality of the marginal costs of signaling, it follows that if a type-$(\iota; g)$ worker is in a discrete pooling set, the other worker pooling with her is $(\hat{\iota}; g)$ defined as:

$$\hat{\iota} = \frac{g}{\alpha} - \iota \equiv \gamma(\iota, g). \tag{3.11}$$

The following lemma will be important for the extension of the model to the GED exam. It links the productivity of two discretely pooled workers with the relative intensity of cognitive skills in the firms' production function.

Lemma 5 *If two workers are discretely pooled, then the less intelligent worker is more productive if the firms' technology is intensive in perseverance ($b < \frac{1}{2}$) and the more intelligent worker is more productive if the firms' technology is intensive in intelligence ($b > \frac{1}{2}$).*

Let $P(\iota, g)$ denote the density of a type-$(\iota; g)$ individual conditional on ι belonging to the pooling-set $\Theta(y(\iota, g), g)$. Then, if ι belongs to a discrete-pooling set, it follows that

$$P(\iota, g) \equiv \frac{p(\iota \mid g)}{p(\iota \mid g) + p(\gamma(\iota) \mid g)}.$$

Furthermore, $P(\iota, g) + P(\gamma(\iota, g), g) = 1$ for all $(\iota; g)$ in a discrete-pooling set.

Analogously to Lemma 3, the local first- and second-order conditions from the workers' incentive-compatibility constraint yield the following:

Lemma 6 *If $(\iota; g)$ belongs to a discrete pooling set, then if an education and wage profile is*

incentive-compatible, they satisfy:

$$y_\iota(\iota, g) = s(\iota, g)\left[P(\iota, g)(bg - \alpha\iota) + P_\iota(\iota, g)\,\iota(g - \alpha\iota)\right] \tag{3.12}$$
$$+ s\left(\frac{g}{\alpha} - \iota, g\right)\left\{[1 - P(\iota, g)]\left[(1-b)g - \alpha\iota\right] + P_\iota(\iota, g)\,\iota(g - \alpha\iota)\right\},$$

$$y_\iota(\iota, g)(g - 2\alpha\iota) \geq 0. \tag{3.13}$$

Equation (3.12) displays how discrete pooling distorts an incentive-compatible profile of education. As in the separated case, equation (3.12) equates the marginal cost with the marginal benefit of education. However, due to the fact that in the discrete pooling case wages are a weighted average of individual worker productivity, the marginal benefit of education in a discrete pooling set is a weighted average of marginal productivities.[13]

In the next subsection, we present some comparative statics results as well as the equilibrium selection criterion.

3.3.4 Equilibrium selection and comparative statics

The proposition below presents some comparative statics results. Since education is costly, a worker would only choose to obtain an additional amount of education if it increases wages. Thus, incentive-compatibility requires wages to be strictly monotonic.

Proposition 20 *In any PBE, wages are strictly increasing and concave in the amount of schooling controlling for the interview.*

Notice that, for fixed ι, the productivity is increasing in the result of the interview g. Then, in a separating set, wages must be increasing in g. However, this may not be true in a pooling set: since wages are a weighed average of the productivity of pooled types (where weights are given by the conditional probability of each type), a change in g would also affect the weights attributed to each type. In a discrete pooling set, for example, it follows that[14]

$$\frac{\partial w}{\partial g} = P(\iota, g)\, f_\eta(\iota, \eta) + [1 - P(\iota, g)]\, f_\eta(\hat{\iota}, \hat{\eta}) + \frac{\partial P(\iota, g)}{\partial g}\left[s(\iota, g) - s(\hat{\iota}, g)\right].$$

[13] Notice that the separating set is a special case of the discrete pooling set where firms are able to infer the workers ability in a pooling set ($P(\iota, g) = 1$).

[14] The same argument also holds for continuous pooling sets.

The first and second terms are positive and represent the direct effect: More productive individuals obtain a higher result in the interview. The last term may be either positive or negative and reflects the indirect effect. If the proportion of more productive individuals is decreasing in g, then this term is negative.[15] If $\iota \mid g$ is uniformly distributed, for example, then this last term vanishes (since the conditional distribution of ι is constant) implying that wages are increasing in the interview.

The difference between the monotonicity of wages in education (Proposition 20) and the non-monotonicity of wages in the interview stems from the fact that education is an endogenous signal while the interview is an exogenous signal. When a costly signal is endogenous, an agent will not purchase an additional amount of it unless she obtains higher wages by doing so. In contrast, when a signal is exogenous, the agent is unable to distort it. Hence, wages may be non-monotonic in this signal.

As the concept of PBE leads to multiple equilibria, we will apply a selection criterion in order to pick an equilibrium. Riley [1979] suggested the concept of a reactive equilibrium that chooses only the separating equilibrium in the continuous-type framework. This concept has been widely applied in the signaling literature.

As a fully separating equilibrium does not exist when the single-crossing property does not hold, one must employ a weaker refinement criterion. We propose the quasi-separability criterion, which consists of a slight modification to the concept of reactive equilibrium (both concepts are equivalent when the SCP holds). Like the reactive equilibrium, the quasi-separability criterion selects the most efficient equilibrium in the class of equilibria with the highest degree of separation.

Definition 6 *A quasi-separable equilibrium is a PBE that satisfies the following conditions:*

1. *If two workers belong to a pooling set, then their marginal cost of schooling must be the same;*

2. *There is no other PBE satisfying condition 1 such that every type obtains less schooling (with strictly less for at least one type).*

[15] Let $s(\iota, g) > s(\hat{\iota}, g)$. Then, $\frac{\partial w}{\partial g} < 0$ if and only if $\frac{\partial P(\iota,g)}{\partial g} < -\frac{P(\iota,g) f_\eta(\iota,\eta)+[1-P(\iota,g)] f_\eta(\hat{\iota},\hat{\eta})}{s(\iota,g)-s(\hat{\iota},g)}$.

The first condition identifies the highest possible degree of separability by ruling out continuous pooling. The second condition gives the boundary condition that uniquely determines the equilibrium. It consists of a Pareto improvement criterion for selection.

The following proposition can be seen as evidence that the SCP does not hold. It states that two individuals with different abilities obtaining the same amount of schooling are not consistent with the SCP. Hence, the fact that the empirical evidence documents that workers with different abilities receive the same wages suggests that the SCP is violated.

Proposition 21 *If there exists pooling in a quasi-separable equilibrium, then the SCP does not hold.*

3.3.5 Characterization of the equilibrium

In this subsection, we characterize the equilibrium of the model. As the results are more technical than the rest of the paper, readers more interested in the applications of the model can skip this subsection.

As in equation (3.11), we denote by $\gamma(\iota, g)$ the type with the same marginal cost of signaling as $(\iota; g)$. We will focus on the case where $\gamma(\iota_0, g) \leq \iota_1$ and $b < 1/2$ (the other cases can be studied in a similar fashion). Clearly, as $\gamma(\iota_0, g) \leq \iota_1$, it follows that $(\gamma(\iota_0, g), \iota_1]$ must be a separating set in any quasi-separable equilibrium (as no other type can have the same marginal cost of schooling as $\iota \in (\gamma(\iota_0, g), \iota_1])$. In this subsection, we show that the quasi-separable equilibrium is such that all types outside of this interval are discretely pooled (where a pool consists of two non-adjacent types). The characterization is carried out through a series of lemmata.

Define the indirect utility $U(\hat{\iota}, \iota, g)$ as the utility received by a type-$(\iota; g)$ worker who gets the contract designed for type $(\hat{\iota}; g)$:

$$U(\hat{\iota}, \iota, g) \equiv w(y(\hat{\iota}, g), g) - c(\iota, g, y(\hat{\iota}, g)).$$

The first lemma establishes another necessary condition for incentive-compatibility.

Lemma 7 *$U(\iota, \cdot, g)$ is continuous for all $\iota \in [\iota_0, \iota_1]$.*

The basic intuition behind this result is that, as the cost of signaling is continuous, if the indirect utility were discontinuous, those individuals in a vicinity of the point of discontinuity could benefit from another type's contract. Hence, it would not be incentive-compatible.

The continuity of U enables us to determine the boundary condition for the amount of education when switching from discrete pooling sets to separating sets. Notice that when a worker becomes pooled with another type, his expected productivity changes discontinuously (as it becomes the average of their productivities). Thus, the wage function has a discontinuity when switching from separating sets to discrete pooling sets. Hence, the education must be discontinuous in order to preserve the continuity of the indirect utility. This is formally established in the following corollary:[16]

Corollary 4 *Let ι be such that $[\iota-\varepsilon,\iota]$ is a discrete pooling set and $(\iota,\iota+\varepsilon]$ is a separating set, for some $\varepsilon>0$. If $y(\cdot,g)$ is right continuous at ι, the following condition is necessary for incentive-compatibility:*

$$y(\iota,g)=\iota(g-\alpha\iota)\left[1-P(\iota,g)\right]\left[s(\iota,g)-s(\gamma(\iota,g),g)\right]+\lim_{x\to\iota_-}y(x,g). \tag{3.14}$$

From now on we assume that the education profile is right continuous. The second lemma determines the maximal discrete pooling set.

Lemma 8 *$[\iota_0,\gamma(\iota_0,g)]$ is the discrete pooling set.*

As the set $(\gamma(\iota_0,g),\iota_1]$ must be separated, it follows from Lemma 8 that the set of types can be partitioned in two intervals: a discrete pooling interval $[\iota_0,\gamma(\iota_0,g)]$ and a separated interval $(\gamma(\iota_0,g),\iota_1]$.

The next lemma determines the boundary condition which gives the equilibrium amount of education. It ensures that the individual with the lowest productivity chooses to get no education.

Lemma 9 *In any quasi-separable equilibrium, $y(\iota_1,g)=0$.*

[16] We obtain the same result if $(\iota,\iota+\varepsilon]$ is a discrete pooling set and $[\iota-\varepsilon,\iota]$ is a separating set for some $\varepsilon>0$.

The proof basically shows that as $(\iota_1; g)$ is separated and is the least productive type, reducing the amount of schooling would never reduce its wages (as no firm would ever expect some type to be less productive than ι_1). But this would also reduce the cost of schooling. Thus, in equilibrium, ι_1 must choose the lowest amount of schooling possible.

The next proposition establishes that the conditions from Lemmata 3, 6, and 9 and Corollary 4 are also sufficient for the quasi-separable equilibrium.

Proposition 22 *A profile of education is a quasi-separable equilibrium if and only if it satisfies the differential equations from Lemma 3 and Lemma 6 and the boundary conditions from Lemma 7 and Lemma 9. Furthermore, the quasi-separable equilibrium exists and is unique*

Proposition 22 is useful as it reduces the problem of obtaining an equilibrium profile of education to that of solving two ordinary differential equations with given boundary conditions.

The amount of education for separated types is determined from condition (1) of Proposition 22 and boundary condition (2). Then, using conditions (3) and (4) (a differential equation with a boundary condition), one can calculate the equilibrium amount of education for discrete pooling types.

Notice that item 4 from Proposition 22 implies that education must jump downward at $\gamma(\iota_0, g)$ since $s(\iota_0, g) - s(\gamma(\iota_0, g), g) > 0$ (see Lemma 5 and $b < 1/2$). This follows from the fact that wages are discontinuous: individuals with $\iota \in \left[\frac{g}{2\alpha}, \gamma(\iota_0, g)\right]$ earn wages higher than their productivity since they are pooled with more productive workers but those with types higher than $\gamma(\iota_0, g)$ earn their productivity since they are separated. Hence, if education were continuous, indirect utility would be discontinuous. But, as shown in Lemma 7, a discontinuous indirect utility is not incentive-compatible. Thus, the amount of education must jump downward in order to preserve the continuity of the indirect utility function.

The main form of the equilibrium can be captured by the following symmetric example. Suppose $\alpha = 1$ and $\Theta = [1, 10]^2$. Then, $g \in [2, 20]$ and

$$c_{y\iota}(\iota, g, y) \left\{ \begin{array}{c} > \\ < \end{array} \right\} 0 \Leftrightarrow \iota \left\{ \begin{array}{c} > \\ < \end{array} \right\} \frac{g}{2}.$$

Consider different values of the interview result g. For the lowest possible value, $g = 2$, the

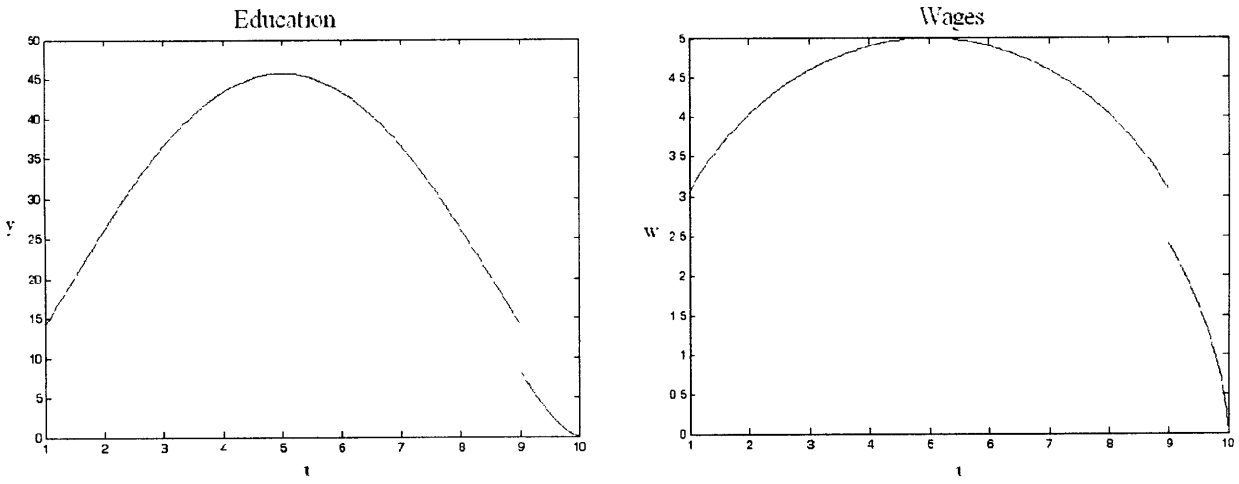

Figure 3-3: Equilibrium in the Symmetric Model (a)

worker's type is uniquely revealed ($\iota = \eta = 1$). Hence, there are no incentives for signaling so that $y(\iota, g) = 0$ for $\iota = 1$, $g = 2$. Analogously, if $g = 20$, the worker's type is uniquely revealed to be $\iota = \eta = 10$ and there are no incentives for signaling.

For $g \in (2, 20)$, the CS_+ and CS_- intervals are $\left[1, \frac{g}{2}\right]$ and $\left[\frac{g}{2}, 10\right]$, respectively. Education is increasing in ι for $\iota \leq \frac{g}{2}$ and decreasing for $\iota \geq \frac{g}{2}$. Due to the symmetry of the example, pairs equidistant from $\frac{g}{2}$ are discretely pooled and extreme types are separated. When $g \in (2, 11)$, the discrete pooling interval is $[1, g-1]$, the separating interval is $(g-1, 10]$ and y is discontinuous at $\iota = g - 1$. When $g \in (11, 20)$, the discrete pooling interval is $[g-10, 10]$, the separating interval is $[1, g-10)$, and y is discontinuous at $\iota = g - 10$.

Figure 3-3 presents the equilibrium amount of education and wages conditional on $g = 10$ for the case where $b = 0.4$, $\alpha = 1$, $\iota_0 = 1$, $\iota_1 = 10$, and $\iota \mid g \sim U[\iota_0, \iota_1]$.[17] Figure 3-4 presents the equilibrium amount of utility and the profile of wages as a function of education conditional on $g = 10$. Notice that both education and wages are discontinuous but the utility is continuous in ι. As Proposition 20 shows, wages are strictly increasing and concave in education.

[17] In a previous version of the paper, we have presented the equilibrium profiles of education, wages, and utility for other parameters. See Araujo, Gottlieb, and Moreira [2007].

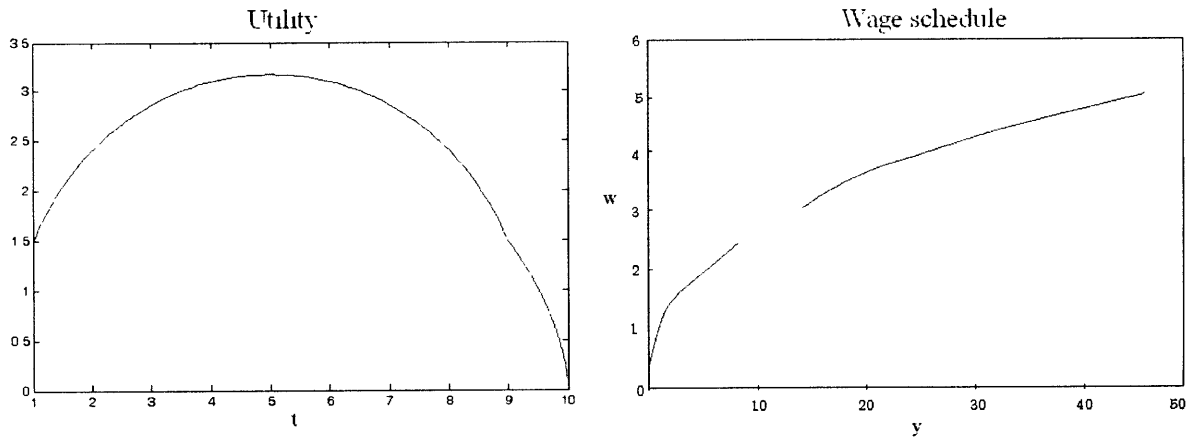

Figure 3-4: Equilibrium in the Symmetric Model (b)

3.4 Countersignaling

In this section, we show how the basic model presented allows us to understand the existence of countersignaling. First, we present a precise definition.

Definition 7 *A type-$(\iota; g)$ worker is countersignaling if*

$$sgn\{y_\iota(\iota, g)\} \neq sgn\{s_\iota(\iota, g)\}.$$

The definition above states that countersignaling occurs if more productive individuals choose less education than intermediate individuals. With no loss of generality, we can restrict our analysis to the case where $b \leq \frac{1}{2}$ (since we can always relabel ι and η).

As shown in Section 3.3.5, education is strictly increasing for $\iota < \frac{g}{2\alpha}$ and strictly decreasing for $\iota > \frac{g}{2\alpha}$. Moreover, as argued in Subsection 3.3.1, the productivity of a worker with interview result g is strictly increasing for $\iota < \frac{bg}{\alpha}$ and strictly decreasing for $\iota > \frac{bg}{\alpha}$. Thus, the countersignaling interval is $\left[\frac{bg}{\alpha}, \frac{g}{2\alpha}\right]$. Hence, countersignaling occurs if, and only if, the schooling technology is different than the firms' technology, i.e., $b \neq \frac{1}{2}$.

Define the distance between the Cobb-Douglas functions $f(\iota, \eta) = \iota^b \eta^{1-b}$ and $\tilde{f}(\iota, \eta) = \iota^{\tilde{b}} \eta^{1-\tilde{b}}$ as $\left|b - \tilde{b}\right|$. Then, the distance from the schooling technology to the firms' technology is given by $\frac{1}{2} - b$. Notice that increasing the distance between the two technologies (i.e., reducing

b) strictly increases the countersignaling interval. Thus, we have proved the following:

Proposition 23 *Countersignaling occurs if and only if the schooling and the firms' technologies are different (i.e., the SCP does not hold), and the countersignaling interval strictly increases in the distance from the schooling technology to the firms' technology.*

This proposition provides an intuitive testable implication. Countersignaling is expected to occur more often in occupations that require a different combination of skills than those required at school. Hence, productive individuals with low levels of education should be more common among sportsmen and artists than among teachers.

The analysis above was made conditional on the interview g. It turns out that education and wages may also be non-monotone in the abilities unconditionally. More specifically, consider the original (two-dimensional) type space. Define the education obtained by type-(ι, η) as

$$\tilde{y}(\iota, \eta) \equiv y(\iota, g(\iota, \eta)).$$

Suppose that $\Theta = [\theta_0, \theta_1]^2$. Then, the worker's type is uniquely revealed when $\iota = \eta = \theta_0$ so that she has no incentive to signal. Analogously, her type is uniquely revealed when $\iota = \eta = \theta_1$. Hence, $\tilde{y}(\theta_0, \theta_0) = \tilde{y}(\theta_1, \theta_1) = 0$ so that $\tilde{y}(\iota, \eta)$ cannot be monotone (as long as $\tilde{y}(\iota, \eta) > 0$ for some ι and η).[18]

The reason for this unconditional non-monotonicity arises from the worker's incentive to signal. When ι and η are extreme, there is not much uncertainty regarding the individual's type. Thus, she faces low incentives to signal. However, when the worker has moderate types, there are many different types with the same interview g. Hence, she has high incentives to signal.[19]

A key message of the model is that, when types are multidimensional, signals that reveal a high type in one dimension may indicate a low type in other dimensions. This result captures the idea that an employer may be suspicious that a potential employee who looks "too perfect" on one dimension may have problems in other unobserved dimensions. One intuitive case is when

[18] When education is productive, $\tilde{y}(\theta_0, \theta_0)$ and $\tilde{y}(\theta_1, \theta_1)$ will generally not be zero.

[19] Therefore, as in Feltovich, Harbaugh, and To [2002], the presence of an additional signal makes intermediate types the ones with the highest incentives to signal.

the (two-dimensional) types are perfectly negatively correlated. Then, one could reparametrize the model into a one-dimensional type model where the SCP does not hold. In the model above, types are perfectly negatively correlated conditional on the interview g. This leads to the SCP not being satisfied and to the emergence of countersignaling.

3.5 The GED exam

3.5.1 Empirical evidence

Signaling models [e.g., Spence, 1973] generally assume that an individual's personal abilities are represented by a scalar of cognitive skills. However a vast body of empirical evidence consistently contradicts this assumption. Heckman, Stixrud, and Urzua [2005], for example, found that for several dimensions of behavior and for a variety of labor market outcomes, non-cognitive skills are better predictors of behavior than cognitive skills.[20]

In the psychology field, the five-factor model of personality (referred to as the "Big Five") identifies five dimensions of non-cognitive characteristics: extroversion, conscientiousness, emotional stability, agreeableness, and openness to experience. Personality measures based on this model are good predictors of job performance for a wide range of professions [Barrick and Mount, 1991].

An interesting set of evidence on the impact of non-cognitive skills on education and wages comes from the General Educational Development (GED). The GED is an exam taken by American high school dropouts to certify that they have equivalent knowledge to high school graduates. It started in 1942 as a way to allow veterans without a high school diploma to obtain a secondary school credential. Today, about half of the students who drop out of high school and a fifth of those classified as "high school graduates" by the U.S. Census Bureau are GED recipients.

[20] Cawley et al. [1996] showed that cognitive ability is only a minor predictor of social performance and that many non-cognitive factors are important determinants of blue collar wages. Bowles and Gintis [2001] provided survey evidence that employers consider measures of non-cognitive skills to be significantly more important than measures of cognitive skills in the hiring of production workers. Klein, Spady and Weiss [1991] showed that lower quit rates and lower absenteeism account for most of the premium awarded by high school graduates compared to high school dropouts (*not* higher productivity). Edwards [1976] showed that dependability and consistency are more valued by blue collar supervisors than cognitive ability and independent thought.

The GED consists of five tests covering mathematics, writing, social studies, science, and literature and arts. Except for the writing section, all the sections consist of multiple choice questions. The costs of acquiring a GED are relatively small. The pecuniary costs range from no cost in some states to around $50 in other states and the median study time for the tests is only about twenty hours.

Even though the U.S. Census classifies dropouts who have acquired a GED as ordinary high school graduates, the market does not treat them equally. GED recipients earn lower wages, work less in any year and stay at jobs for shorter periods than high school graduates [Boesel, Alsalam and Smith, 1998].

GED recipients are smarter than other dropouts (as measured by IQ) but exhibit more behavior and self discipline problems and are less able to finish tasks. They switch jobs at a faster rate and are more likely to skip school, fight at school and work, use marijuana, and participate in robberies. Hence, the GED conveys two pieces of information in one signal. The student who acquires it is bright, but lacks perseverance and self discipline [Cameron and Heckman, 1993; Cavallo, Heckman and Hsee, 1998; and Heckman and Rubinstein, 2001].

Cavallo, Heckman and Hsee [1998] and Heckman and Rubinstein [2001] have shown that when one controls for both cognitive and non-cognitive abilities, there is no difference in earnings between a GED recipient and a dropout who has not acquired the certificate. Tyler, Murnane, and Willett [2000] obtained similar results except for young white dropouts who were in the margin of passing the exam. As for females, the evidence is the same as that of males, except for those who dropped out because of pregnancy [Carneiro and Heckman, 2003].

Because high school dropouts who have taken the GED are treated in the labor market just like those who have not taken it, any theory that tries to explain this exam must exhibit pooling in equilibrium. Moreover, since GED recipients do not earn higher wages than dropouts without the GED, the signal-earnings relation is not strictly monotone as in the traditional signaling models. As Heckman, Stixrud, and Urzua [2005] point out,

> Our evidence that multiple abilities determine schooling challenges the conventional single skill signalling model due to Arrow (1973) and Spence (1973). A special challenge is the GED program where the credential (the GED test) conveys multiple conflicting signals. GED recipients are smarter than other high school dropouts but

they have lower noncognitive skills.

3.5.2 The Model

In this subsection, we extend the basic framework to study the effect of the introduction of a pass-or-fail test like the GED. We model the GED as a certifiable statement that only individuals with a sufficiently high combination of characteristics are able to reveal. Hence, we will add a disclosure dimension to the signaling model presented previously.[21]

We model the GED as an additional signal $h(\iota, \eta)$ that only individuals with a sufficiently high combination of characteristics are able to receive. More specifically, denoting by $h(\iota, \eta) = 1$ if an individual passes the exam and $h(\iota, \eta) = 0$ if she fails, we specify the test as

$$h = \begin{cases} 1, \text{ if } \kappa\iota + \eta \geq \overline{g} \\ 0, \text{ if otherwise} \end{cases},$$

where $\overline{g} \in \mathbb{R}_{++}$ is the parameter that represents the minimum combination of skills required to pass the test (passing standards) and κ is the rate of substitution between intelligence and perseverance.[22]

We assume that there is a nonnegative expected net benefit in passing the test. The positive expected net benefit may be due to three characteristics of the GED. First, the costs of taking the exam are rather low. The median time studying for the GED is 20 hours and the monetary costs range from \$0 to \$50. Second, the GED provides recipients with the option of postsecondary schooling and joining the U.S. Military. Third, there could be some nonmonetary benefits for being legally considered a high-school graduate.

After obtaining the GED, the worker may choose whether or not to disclose this information to the employer. Hence, firms cannot distinguish between workers who have acquired a GED but chose not to disclose it from those who were unable to obtain the GED. We denote by $\tilde{h}$ the information disclosed by the worker. More specifically, $\tilde{h} = 1$ if $h = 1$ and the worker chooses

[21] See, for example, Grossman (1981) and Okuno-Fujiwara, Postlewaite, and Suzumura (1990) for standard disclosure games.

[22] The assumption that schooling does not affect the possibility of passing the GED is unimportant for our results. As would probably be clear, all results still hold if education entered linearly in the minimum combination of skills.

to disclose this information. Otherwise, $\tilde{h} = 0$. Firms observe $\tilde{h}$ but not h. Therefore, employers now observe the amount of education y, the interview result g, and the GED $\tilde{h}$.

Controlling for the interview result g, h can be rewritten as

$$h = \begin{cases} 1, \text{ if } (\kappa - \alpha)\iota \geq \overline{g} - g \\ 0, \text{ otherwise.} \end{cases}.$$

Since the GED exam is intensive in cognitive skills, we shall assume that the exam h emphasizes intelligence more than the interview g does:

Assumption 1 $\kappa > \alpha$.

Then, each worker with $\iota \geq \frac{\overline{g}-g}{\kappa-\alpha}$ would be able to pass the test. The graphs in Figure 5 separate the interval $[\iota_0, \iota_1]$ in three regions. The first graph depicts the case where $\frac{\overline{g}-g}{\kappa-\alpha} > \frac{g}{2\alpha}$, while the second graph represents the case where $\frac{\overline{g}-g}{\kappa-\alpha} < \frac{g}{2\alpha}$.

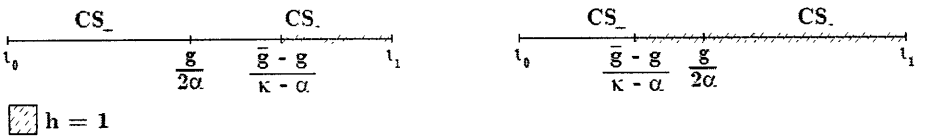

Figure 3-5: CS_+ and CS_- regions

In the left region, workers have low intelligence so that education must be increasing in intelligence (CS_+ region) and the worker is unable to pass the test. On the right side, workers have high intelligence. Thus, education must be decreasing in intelligence (CS_- region) and the worker is able to pass the test.

The region in the middle depends on the sign of $\frac{\overline{g}-g}{\kappa-\alpha} - \frac{g}{2\alpha}$. If $\frac{\overline{g}-g}{\kappa-\alpha} > \frac{g}{2\alpha}$ (first graph), some workers with types in the CS_- region are unable to receive $h = 1$. If $\frac{\overline{g}-g}{\kappa-\alpha} < \frac{g}{2\alpha}$ (second graph), some workers with types in the CS_+ region are able to pass the test.

The game consists of adding a last stage to the game presented in Section 3.2. Hence, the timing of the game is as follows. First, nature determines each worker's type according to density

p. Second, workers choose their education $y^g(\iota, g)$ contingent on their types.[23] Subsequently, firms offer a wage $w^g\left(y, g, \tilde{h}\right)$ conditional on observing (y, g) and on whether the worker will $(\tilde{h} = 1)$ or will not disclose the GED $(\tilde{h} = 0)$. Then, workers choose whether or not to acquire the GED and, if they do, whether or not to disclose this information.[24]

Consider the case where the firms' technology is intensive in non-cognitive skills. Then, for any pool of workers, the one with higher perseverance/lower intelligence is the most productive. If firms could identify the most intelligent individuals in a pool of workers, they would offer them lower wages. But, of course, a worker would never take the GED if this reduced her earnings. Hence, in the case where the firms' technology is intensive in non-cognitive skills, allowing workers to take the GED does not affect education and wages (compared with the equilibrium obtained in Section 3.3). Thus, we say that, in this case, the GED is a neutral signal.[25] This conclusion, which is the main result of this section, is formally stated in the following proposition.

Proposition 24 *Suppose that the firms' technology is intensive in non-cognitive skills. Let* $\{y(\iota, g),\ w(y(\iota, g), g)\}$ *denote the quasi-separable equilibrium of the model where the GED is not available. There exists a quasi-separable equilibrium of the game where the GED is available such that acquiring the GED does not affect education and wages:*

$$\begin{aligned} y^g\left(\iota, g, \tilde{h}^*(\iota, g)\right) &= y(\iota, g), \\ w^g\left(y(\iota, g), g, \tilde{h}^*(\iota, g)\right) &= w(y(\iota, g), g), \\ \tilde{h}^*(\iota, g) &= 0 \quad \forall \iota \in [\iota_0, \gamma(\iota_0)], \\ h^*(\iota, g) &= 1 \iff \iota \geq \frac{\overline{g} - g}{\kappa - \alpha}. \end{aligned} \tag{3.15}$$

[23] We add the superscript g to differentiate the model where the GED is available from the one examined in the previously.

[24] In standard disclosure models, individuals first choose which information to reveal A second stage consisting of a signaling game is easily introduced by assuming that the payoffs from the disclosure game are obtained by backward induction from the second stage. In the model presented above, however, the signaling game occurs in the first stage and the disclosure game occurs in the second stage. This is the natural assumption in our model since the choice of whether or not to take the GED is usually made after the worker has decided how much education to acquire.

[25] When ι and $\gamma(\iota)$ are both able to obtain the GED (i.e., they are greater than $\frac{\overline{g}-g}{\kappa-\alpha}$), there also exist equilibria such that $\tilde{h}^*(\iota, g) = \tilde{h}^*(\gamma(\iota), g) = 1$. Since the GED does not disclose any information in this case, it does not affect education or wages.

Proof. The result is trivial for a separating set. Assume two workers with $\iota \leq \frac{\overline{g}-g}{\kappa-\alpha} \leq \hat{\iota}$ are pooled in the same contract (otherwise, the signal is not informational). Since acquiring the GED has positive net benefits and a worker can always choose not to disclose that she has obtained the GED, $\hat{\iota}$ will have $h = 1$ and ι will have $h = 0$. Suppose that type-$\hat{\iota}$ chooses to disclose this information – i.e. $\tilde{h}(\hat{\iota}, g) = 1$. Then, from Lemma 5, the firm would offer a lower salary for the type-$\hat{\iota}$ worker and a higher salary for ι. But this cannot be an equilibrium since the type-$\hat{\iota}$ worker's strategy is not optimal (condition 1 from Definition 4). Thus, all types who were discretely pooled in the quasi-separable equilibrium of the game where the GED was not available choose $\tilde{h} = 0$. Then, it follows that the separating and discrete pooling intervals as well as the conditions for the PBE obtained in Proposition 22 are the same in both games, which concludes the proof. ■

Notice that, consistent with Heckman and Rubinstein [2001] and Cavallo, Heckman and Hsee [1998], workers who have a GED have higher cognitive skills and lower non-cognitive skills but receive the same wages as those who do not have it. However, as the result above holds for all $\overline{g} \in \mathbb{R}_{++}$, it follows that, unlike Cavallo, Heckman and Hsee [1998] suggested, an increase in the GED standards, $\overline{g}$, would not affect the equilibrium education and wage schedules.[26] Furthermore, since the introduction of the GED does not affect the equilibrium profile of education, our model does not support the claim that, when the GED is neutral, it may discourage education [e.g., Cavallo, Heckman and Hsee, 1998].

A key assumption for the neutrality of the GED is that the firms' technology is intensive in non-cognitive abilities.[27] Consider now the case where the firms' technology is intensive in cognitive skills. Then, for any pool of workers, the one who is able to acquire the GED is the most productive. Hence, by disclosing that one has a GED, a worker is able to obtain higher wages at no cost so that the GED is no longer neutral. The next proposition formally proves this result.

Proposition 25 *Suppose that the firms' technology is intensive in cognitive skills and* $\frac{\overline{g}-g}{\kappa-\alpha} \in$

[26] This implication of the model could be tested as passing standards vary by states and have changed over time. Thus, one could test if the neutrality of the GED is robust to different states and different periods of time.

[27] The neutrality of the GED does not rely on the assumption that education does not affect the ability to pass the GED exam. For example, suppose that an individual would be able to pass on the GED if $\kappa\iota + \eta + \beta y \geq \overline{g}$. Then, the shaded area in Figure 8 would depend on y but if two workers were discretely pooled in a contract, the one who could pass the test would still be the least productive worker.

$[\iota_0, \gamma(\iota_0)]$. *Then, there exists an equilibrium where the GED is non-neutral: for any* $\iota \leq \frac{\overline{g}-g}{\kappa-\alpha} \leq \gamma(\iota)$,

$$w^g\left(y(\iota,g),g,\tilde{h}^*(\iota,g)\right) < w(y(\iota,g),g) < w\left(y(\gamma(\iota),g),g,\tilde{h}^*(\gamma(\iota),g)\right),$$
$$\tilde{h}^*(\iota,g)=0, \quad \tilde{h}^*(\gamma(\iota),g)=1,$$
$$h^*(\iota,g)=1 \iff \iota \geq \frac{\overline{g}-g}{\kappa-\alpha}.$$

Proof. In this case, condition 2 of Definition 4 implies that $w^g(y(\gamma(\iota),g),g,1) = s(\gamma(\iota)) > s(\iota) = w^g(y(\iota,g),g,0)$, where the inequality follows from Lemma 5. Then, type $\gamma(\iota)$ prefers to disclose $\tilde{h}=1$ and type ι cannot pool with her since she cannot acquire the GED. The existence part follows the same steps as Section 3.3, except that now there will be separability in the two extremes. ■

Corollary 5 *Suppose that the firms' technology is intensive in non-cognitive skills. There exists an equilibrium such that an exam* h *that places more weight to non-cognitive skills* ($\kappa < \alpha$) *is non-neutral.*

A way to make the GED exam a non-neutral signal would be to put more emphasis on non-cognitive skills as it would separate two pooled workers with different signs, h. Even though it must be significantly harder to design a signal that emphasizes non-cognitive skills, psychologists have developed tests that measure such skills, and they have been used by companies to screen workers [e.g., Sternberg, 1985].

When the GED is non-neutral ($b > \frac{1}{2}$), it separates two previously pooled workers. Then, the wage received by the more (less) productive worker increases (decreases). As incentive-compatibility requires that the indirect utility must be continuous, it follows that, in this case, the introduction of the GED increases (decreases) the education obtained by the more (less) productive workers. Hence, another testable implication of the model is that the variance of education (conditional on g) should increase when the GED is non-neutral and should remain constant when it is neutral.

3.6 Other Applications

In the previous sections, we have shown that a two-dimensional signaling game featuring an additional exogenous signal can be reduced to a one-dimensional signaling game where the SCP is violated. Moreover, we have characterized the equilibrium of this one-dimensional signaling game.

Although the paper is presented in a job market environment, it can be employed in a wide variety of environments. In this section, we briefly discuss some examples.

First, consider a corporate finance context, where firms may use dividends in order to signal future earnings. Reinterpret ι as current earnings, η as future earnings, y as the amount of dividends paid, and g as representing a specialist's classification of the profitability of the firm. Then, if the firm's and the specialist's time preferences were not aligned, the SCP would not hold in general (see Appendix A). This misalignment might be due to credit constraints. It could also be a consequence of an (unmodeled) CEO remuneration contract that induces greater short-term orientation.[28] Consistently with Benartzi, Michaely, and Thaler [1997], our model would then predict a non-monotonic relation between dividends and future earnings. Furthermore, higher dividends would be a mixed signal. Depending on the time preference of the firm and investors, it could signal both high present earnings and low future earnings or low present earnings and high future earnings.

In an international finance context, we could reinterpret ι as the government's commitment to maintaining a fixed exchange rate, η as the quality of the fundamentals of the country, and y as the interest rate. The signal g could denote the country's risk classification or some other indicator of its fundamentals. As long as the risk classification is not perfectly aligned with the government's preferences in the sense that their marginal rates of substitution between ι and η cannot be ordered, the SCP would not hold. In that case, consistently with the evidence from Drazen and Hubrich [2003], our model would predict that interest rates are mixed signals as they indicate that the government is committed to maintaining a fixed exchange rate, but may also signal weak fundamentals. Furthermore, countersignaling implies that it could be optimal for a country to choose lower interest rates in order to signal strong fundamentals.

[28] Bolton, Scheinkman, and Xiong [2005] have shown that when the equilibrium stock prices may differ from the fundamental value, the optimal contract may induce greater short-term orientation.

Another application is a model where firms choose the amount of advertising expenditures in order to signal product quality, captured by a two-dimensional type (ι, η). Reinterpreting y as advertising expenditures, g as additional information obtained from other sources (e.g., word-of-mouth advertising, magazine reviews, etc.), one can easily apply the model presented above. In that model, advertising may be a mixed signal. Furthermore, consistent with the evidence from Caves and Greene [1996], Clements [2004], and Orzach et al. [2002], there would not be an increasing relationship between advertising and product quality.

3.7 Conclusion

In this chapter, we presented a model of mixed signals. We demonstrated that when firms have access to an interview technology, the two-dimensional model can be reduced to a one-dimensional model where the single-crossing property may not hold. When this is the case, the equilibrium features countersignaling in the sense that signals are non-monotone in the worker's productivity.

It was shown that countersignaling occurs if, and only if, the schooling technology differs from the firm's technology. Moreover, the countersignaling interval is strictly increasing in the distance between the schooling and the firm's technologies. Hence, this phenomenon is expected to be more important in occupations that require a more diverse combination of skills from those required in the schooling process.

We have extended the basic model in order to analyze the GED exam. It was shown that, consistently with the empirical evidence, a GED recipient has above average cognitive skills and below average non-cognitive skills. When cognitive skills are more valued in the labor market, this new information affects equilibrium wages. However, when non-cognitive skills are more valued in the labor market than cognitive skills (as suggested by significant empirical evidence), it does not affect the wage schedule.

The main problem with the GED is its focus on cognitive skills. As the firm's main concern is usually about the worker's non-cognitive skills, a non-neutral signal should assign more weight to this kind of skills. Thus, changing its focus to non-cognitive skills would turn the GED into a non-neutral signal. Moreover, increasing the passing standards with no change of the relative

intensity of each skill in the test would not change the equilibrium wages.

Our results provide evidence of the importance of the failure of the single-crossing property in explaining observed phenomena. As the absence of this property is necessary for the existence of discrete pooling in equilibrium, the fact that an individual with high cognitive ability and low non-cognitive ability receives the same wages as another with low cognitive ability and high non-cognitive ability while an individual with intermediate abilities does not is evidence of lack of the single-crossing property.

This essay also has a technical contribution as it characterizes the equilibrium of a signaling model where the single-crossing condition does not hold. This framework can be employed in a wide variety of environments such as advertising, corporate finance, and international finance.

Appendix A Robustness of the Single-Crossing Property

This section characterizes the set of functions c and g that satisfy for the single-crossing property (SCP). We will show that the SCP does not hold as long as the interview technology and the schooling technology cannot be ordered according to their technical rates of substitution.

Let the cost of signaling be represented by the twice continuously differentiable function

$$c = \frac{y}{r(\iota, \eta)},$$

which is assumed to be strictly decreasing in ι and η and strictly increasing in y.

The interview technology is represented by the twice continuously differentiable function $g(\iota, \eta)$ which is assumed to be strictly increasing. From the implicit function theorem, there exists $\varphi(\iota, \bar{g})$ such that

$$\varphi(\iota, \bar{g}) = \eta$$

if and only if $g(\iota, \eta) = \bar{g}$. Moreover,

$$\varphi_\iota = -\frac{g_\iota}{g_\eta}.$$

Substituting into the cost function, it follows that the cost of signaling function can be written as $c = \frac{y}{r(\iota, \varphi(\iota, \bar{g}))}$.Hence,

$$c_{y\iota} = -\frac{r_\iota - r_\eta \times \frac{g_\iota}{g_\eta}}{\left[r(\iota, \varphi(\iota, \bar{g}))\right]^2}.$$

Thus, the SCP holds if, and only if, $\frac{r_\iota}{r_\eta} - \frac{g_\iota}{g_\eta}$ has a constant sign for all ι, η. Therefore, a necessary

and sufficient condition for the SCP to hold is that the technical rates of substitution of the schooling technology and the interview technology can be ordered.

Suppose, for example, that w and g are both CES functions:[29]

$$r = [\alpha_1 \iota^\rho + \alpha_2 \eta^\rho]^{\frac{1}{\rho}},$$
$$g = [\beta_1 \iota^\gamma + \beta_2 \eta^\gamma]^{\frac{1}{\gamma}}.$$

Then, the SCP holds if, and only if, $\frac{\eta}{\iota} - \left(\frac{\beta_1}{\alpha_1}\frac{\alpha_2}{\beta_2}\right)^{\frac{1}{\gamma-\rho}}$ has a constant sign for all ι, η.

When the SCP does not hold and the CS_+ and CS_- regions are not independent of y, the necessary conditions may no longer be sufficient and we cannot guarantee that an equilibrium exists. If it exists, however, some results presented in this paper still hold. As in Remark 9, it can be shown that we cannot in general have a fully separating equilibrium when the SCP is violated. The equilibrium must be such that y is increasing in the CS_+ region and decreasing in the CS_- region so that, as long as it does not feature complete pooling, y will not be monotone.

In a previous version of the paper, we have analyzed the case where education affects the interview result so that the CS_+ and CS_- regions are no longer independent of y. We have shown that if a quasi-separable equilibrium exists, then our results on countersignaling and the GED still hold.[30]

Appendix B Proofs

Proof of Lemma 3:

Since the first claim of this lemma is a particular case of Lemma *6*, we will only prove Lemma *6*. The second claim follows from equation (3.7), since

$$y_\iota(\iota, g) > 0 \Longleftrightarrow \iota < \frac{bg}{\alpha} = \iota^*(g). \tag{3.16}$$

■

Proof of Lemma 4:

[29] The functions considered in the model are special cases of the CES when $\gamma = 1$, $\beta_1 = \alpha$, $\beta_2 = 1$, $\rho \to 0$, and $\alpha_1 = \alpha_2 = 1$.

[30] See Araujo, Gottlieb, and Moreira [2007].

If $\{w(y(\iota,g)),y(\iota,g)\}$ is an incentive-compatible profile of education and wages, it must satisfy

$$\iota \in \arg\max_{\tilde{\iota}} w(y(\tilde{\iota},g),g) - c(\iota,g,y(\tilde{\iota},g)).$$

The first-order condition of the program above yields

$$w_y(y(\iota,g),g) = c_y(\iota,g,y(\iota,g)). \tag{3.17}$$

Suppose that $y(\iota,g) = y(\tilde{\iota},g)$ for some regular types ι and $\tilde{\iota}$. Substituting in equation (3.17) yields $c_y(\iota,g,y(\iota,g)) = c_y(\tilde{\iota},g,y(\tilde{\iota},g))$. ■

Proof of Lemma 5:

Let $\iota > \hat{\iota}$ be two discretely pooled workers and notice that $\alpha\hat{\iota} = \eta$ and $\alpha\iota = \hat{\eta}$. Substituting in the firm's technology yields,

$$f(\iota,g) > f(\hat{\iota},g) \Longleftrightarrow \iota^b\hat{\iota}^{1-b} > \hat{\iota}^b\iota^{1-b} \Longleftrightarrow 2b > 1.$$

■

Proof of Lemma 6:

Define $U(\hat{\iota},\iota)$ as the expected utility received by a type-(ι,g) individual who gets a contract designed for type $(\hat{\iota},g)$:

$$U(\hat{\iota},\iota) = P(\hat{\iota},g)\,s(\hat{\iota},g) + P\left(\frac{g}{\alpha}-\hat{\iota},g\right)s\left(\frac{g}{\alpha}-\hat{\iota},g\right) - c(\iota,g,y(\hat{\iota},g)).$$

The incentive-compatibility constraint is

$$\iota \in \arg\max_{\hat{\iota}\in[\iota_0,\iota_1]} U(\hat{\iota},\iota), \qquad \forall\iota \in [\iota_0,\iota_1].$$

The local first-order condition is

$$U_{\hat{\iota}}(\hat{\iota},\iota)|_{\hat{\iota}=\iota} = 0, \qquad \forall\iota \in [\iota_0,\iota_1]. \tag{3.18}$$

Calculating the derivative above yields

$$c_y(\iota, g, y(\iota, g))\, y_\iota(\iota, g) = P_\iota(\iota, g)\, s(\iota, g) + P(\iota, g)\, s_\iota(\iota, g) - P_\iota\left(\frac{g}{\alpha} - \iota, g\right) s\left(\frac{g}{\alpha} - \iota, g\right) - P\left(\frac{g}{\alpha} - \iota, g\right) s_\iota\left(\frac{g}{\alpha} - \iota, g\right).$$

Since $s_\iota(x, g) = \frac{bg - \alpha x}{x(g - \alpha x)} s(x, g)$ and $c_y(\iota, g, y(\iota, g)) = \frac{1}{\iota(g - \alpha\iota)}$, the expression becomes

$$y_\iota(\iota, g) = s(\iota, g)\left[P_\iota(\iota, g)\, \iota(g - \alpha\iota) + P(\iota, g)(bg - \alpha\iota)\right] + s\left(\frac{g}{\alpha} - \iota, g\right)\left[P\left(\frac{g}{\alpha} - \iota, g\right)[g(1 - b) - \alpha\iota] - P_\iota\left(\frac{g}{\alpha} - \iota, g\right)\iota(g - \alpha\iota)\right].$$

Using the fact that $P\left(\frac{g}{\alpha} - \iota, g\right) = 1 - P(\iota, g)$ for all ι, we obtain equation (3.12).

Differentiating equation (3.18) with respect to ι, we obtain

$$U_{\tilde{\iota}\tilde{\iota}}(\iota, \iota) + U_{\tilde{\iota}\iota}(\iota, \iota) = 0. \tag{3.19}$$

The necessary second-order condition is

$$U_{\tilde{\iota}\tilde{\iota}}(\iota, \iota) \leq 0. \tag{3.20}$$

Substituting (3.19) in (3.20), it follows that

$$U_{\tilde{\iota}\iota}(\iota, \iota) = -c_{y\iota}(\iota, g, y(\iota, g))\, y_\iota(\iota, g) \geq 0.$$

Substituting $c_{y\iota}(\iota, g, y) = -\frac{g - 2\alpha\iota}{[\iota(g - \alpha\iota)]^2}$ in the inequality above establishes (3.13). ∎

Proof of Proposition 20:

Suppose that wages are not strictly increasing in education. Then, there exist types ι and $\tilde{\iota}$ such that

$$y(\iota, g) > y(\tilde{\iota}, g) \text{ and } w(y(\iota, g), g) \leq w(y(\tilde{\iota}, g), g).$$

But this is not incentive-compatible since

$$w(y(\iota, g), g) - \frac{y(\iota, g)}{\iota\eta} < w(y(\tilde{\iota}, g), g) - \frac{y(\tilde{\iota}, g)}{\iota\eta},$$

concluding the first part of the proof.

In order to establish the concavity of w, consider the incentive-compatibility constraint:

$$y(\iota, g) \in \arg\max_{y} w(y, g) - \frac{y}{\iota(g - \alpha\iota)}.$$

The second-order condition (necessary) is[31]

$$w_{yy}(y(\iota, g), g) \leq 0.$$

■

Proof of Proposition 21:

Suppose that type ι belongs to a pooling set. Then, there exists a type $\hat{\iota} = \frac{g}{\alpha} - \iota \neq \iota$ that pools in a contract with ι. Hence, $\iota + \hat{\iota} = \frac{g}{2\alpha}$, implying that ι and $\hat{\iota}$ cannot both belong to CS_+ or CS_-. ■

Proof of Lemma 7:

Suppose that ι is an interior point of either a separating set or a discrete pooling set. Then, as y is continuous in ι (since it solves a differential equation) it follows that:

$$\lim_{x \to \iota_-} U(\iota, x) = \lim_{x \to \iota_+} U(\iota, x) = U(\iota, \iota).$$

Suppose that $[\iota - \varepsilon, \iota)$ is a discrete pooling set and $[\iota, \iota + \varepsilon]$ is a separating set, for some $\varepsilon > 0$. Clearly, a necessary condition for incentive-compatibility is

$$\lim_{x \to \iota_+} U(x, x) \geq \lim_{x \to \iota_-} U(\iota, x),$$

which means that the first individuals in the separating set would not want to get the contract of the last individual in the discrete pooling set. Then,

$$\lim_{x \to \iota_+} U(x, x) = s(\iota, g) - \frac{y(\iota, g)}{\iota(g - \alpha\iota)},$$

$$\lim_{x \to \iota_-} U(\iota, x) = P(\iota, g)\, s(\iota, g) + [1 - P(\iota, g)]\, s(\gamma(\iota, g), g) - \frac{\lim_{x \to \iota_-} y(x, g)}{\iota(g - \alpha\iota)}$$

since $y(\cdot, g)$ is right continuous at ι.

[31] Another way of establishing the monotonicity of w consists of calculating the first-order condition of the indirect mechanism, which yields: $w_y(y(\iota, g), g) = \frac{1}{\iota(g - \alpha\iota)} > 0$.

Thus, the inequality can be written as

$$y(\iota, g) \leq \lim_{x \to \iota_-} y(x, g) + \iota(g - \alpha\iota)[1 - P(\iota, g)][s(\iota, g) - s(\gamma(\iota, g), g)].$$

Another necessary condition for incentive-compatibility is

$$\lim_{x \to \iota_-} U(x, x) \geq \lim_{x \to \iota_+} U(x, \iota),$$

which states that the last individuals in the discrete pooling set would not want to get the contract of the first individuals in the separating set.

Using the definition of the indirect utility, we get

$$\lim_{x \to \iota_-} U(x, x) = P(\iota, g)\, s(\iota, g) + [1 - P(\iota, g)]\, s(\gamma(\iota, g), g) - \frac{\lim_{x \to \iota_-} y(x, g)}{\iota(g - \alpha\iota)},$$
$$\lim_{x \to \iota_+} U(x, \iota) = s(\iota, g) - \frac{y(\iota, g)}{\iota(g - \alpha\iota)},$$

implying that

$$y(\iota, g) \geq \lim_{x \to \iota_-} y(x, g) + \iota(g - \alpha\iota)[1 - P(\iota, g)][s(\iota, g) - s(\gamma(\iota, g), g)].$$

Combining these two necessary conditions, we obtain:

$$y(\iota, g) = \frac{\iota(g - \alpha\iota)[s(\iota, g) - s(\gamma(\iota, g), g)]}{2} + \lim_{x \to \iota_-} y(x, g). \tag{3.21}$$

Substituting in the indirect utility function, it follows that $U(\iota, \iota) = \lim_{x \to \iota} U(x, \iota)$. ∎

Proof of Lemma 8:

From Remark 10, it follows that some types between $\frac{bg}{\alpha}$ and $\frac{g}{2\alpha}$ must be discretely pooled (since there is no continuous pooling in a quasi-separable equilibrium). Assume that some type in $[\iota_0, \gamma(\iota_0)]$ is separated. Then, there must be a $\iota \in \left[\iota_0, \frac{g}{2\alpha}\right]$ such that $\left[\iota, \frac{g}{2\alpha}\right]$ is a discrete pooling set and $[\iota - \varepsilon, \iota)$ is a separated set for some $\varepsilon > 0$. From equation 7, it follows that $y(\iota, g) < \lim_{x \to \iota_{0-}} y(x, g)$ (i.e., y jumps upward when the types become separated). But this is not incentive-compatible because the marginal cost of education is lower for $\iota + \varepsilon$ than for $\iota - \varepsilon$ for ε sufficiently small (thus, a type-$(\iota + \varepsilon)$ individual would always prefer to get the type-$(\iota - \varepsilon)$ individual's contract). ∎

Proof of Lemma 9:

As $\gamma(\iota_1, g) < \iota_0$, ι_1 is separated. Suppose a type ι_1 worker chooses some strictly positive education $\tilde{y} > 0$. Then, according to equation (3.4), this worker's wage must be $s(\iota_1, g)$ in any separating equilibrium (which is the lowest wage since ι_1 is the least productive type). However, she would receive a wage of at least $s(\iota_1, g)$ if she chose $y = 0$. As $y = 0$ implies in a lower signaling cost and does not reduce her utility, she would be strictly better off by doing so. ∎

Proof of Proposition 22:

Let where $y(\iota, g)$ be given by the solution to the differential equations from Lemma 3 and Lemma 6 with the boundary conditions from Lemma 7 and Lemma 9. Define $w(y, g)$ as in condition 2 from Definition 4. Let μ be a Dirac measure concentrated at ι in the interval $[\gamma(\iota_0, g), \iota_1]$ and $P(\iota, g)$ in the interval $[\iota_0, \gamma(\iota_0, g))$ for y in the range of signals.[32]

By construction, in order to show that $\{y(\iota, g), w(y, g)\}$ and $\mu(.|y, g)$ is a PBE, it suffices to establish that

$$y(\iota, g) \in \arg\max_y w(y, g) - c(\iota, g, y)$$

for all $\iota \in [\iota_0, \iota_1]$.

First, observe that the first-order condition of the program above is equivalent to equation (3.7) for $\iota \in [\gamma(\iota_0, g), \iota_1]$ and equation (3.12) for $\iota \in [\iota_0, \gamma(\iota_0, g))$ and, therefore, are satisfied by $y(\iota, g)$. Moreover, the (global) second-order condition is equivalent to

$$w_{yy}(y, g) - c_{yy}(\iota, g, y) \leq 0$$

for y in the range of signals.

From equation (3.1), $c_{yy}(\iota, g, y) = 0$. Then, since $w_y(y, g) = c_y(\iota, g, y)$ for $y = y(\iota, g)$ by the first-order condition,

$$w_{yy}(y, g) y_\iota = c_{yy}(\iota, g, y) y_\iota + c_{y\iota}(\iota, g, y)$$

for $y = y(\iota, g)$. Thus, whenever $y_\iota \neq 0$.

$$w_{yy}(y, g) = \frac{c_{y\iota}(\iota, g, y)}{y_\iota(\iota, g)} \leq 0$$

for $y = y(\iota, g)$, by equations (8) and (14). Therefore, the (global) second-order condition holds.

To complete the first part of the proof we have to show that this PBE is the quasi-separable equilibrium. But this is clear by inspection of Definition 3 and Lemma 7. Existence and uniqueness of the

[32] For y outside the range of signals, let μ be a Dirac measure concentrated at ι_1 (which is the least productive type by Lemma 3).

quasi-separable equilibrium follow from the fact that both differential equations are Lipschitz. ∎

References

Araujo, A., Gottlieb, D., and Moreira, H. (2007). "A model of mixed signals with applications to countersignaling and the GED", Working Paper, Getulio Vargas Foundation.

Araujo, A., Gottlieb, D., and Moreira, H. (2007). "A model of mixed signals with applications to countersignaling", *RAND Journal of Economics,* **38**, 1020-1043.

Araujo, A. and Moreira, H. (2001). "Adverse Selection Problems without the Spence-Mirrlees Condition", Working Paper no. 425, Getulio Vargas Foundation.

Arrow, K. J. (1973) "Higher education as a filter", *Journal of Public Economics,* **2**, 193-216.

Barrick, M. and Mount, M. (1991). "The Big Five personality dimensions and job performance: a meta-analysis", *Personnel Psychology*, **44**, 1-26.

Benabou, R. and Tirole, J. (2006). "Incentives and Prosocial Behavior", *American Economic Review*, **96**, 1652-1678.

Benartzi, S., Michaely, R. and Thaler, R. H. (1997). "Do Changes in Dividends Signal the Future or the Past", *Journal of Finance*, **52**, 1007-1034.

Boesel, D., Alsalam, N. and Smith, T. M. (1998). "Educational and Labor Market Performance of GED Recipients", Washington, DC: National Library of Education, Office of Educational Research and Improvement, U.S. Department of Education.

Bolton, P., Scheinkman, J. A., and Xiong, W. (2005). "Executive Compensation and Short-termist Behavior in Speculative Markets", *Review of Economic Studies*, forthcoming.

Bowles, S. and Gintis, H. (2001). "Schooling in Capitalist America Revisited", *Sociology of Education*, **75**, 1-18.

Cameron, S. and Heckman, J. (1993). "The Nonequivalence of High School Equivalents", *Journal of Labor Economics*, January, **11**, 1-47.

Carneiro, P. and Heckman, J. (2003). "Human Capital Policy", Woking Paper no. 9495, National Bureau of Economic Research.

Cavallo, A., Heckman, J., and Hsee, J. (1998) "The GED is a mixed signal", Unpublished manuscript presented at the meetings of the American Economic Association, 3 January, New York.

Caves, R.E. and Greene, D. P. (1996). "Brands' Quality Levels, Prices, and Advertising Outlays: Empirical Evidence on Signals and Information Costs", *International Journal of Industrial Organization*, **14**, 29-52.

Cawley, J., Conneely, K., Heckman, J., and Vytlacil, E. (1996). "Measuring the Effects of Cognitive Ability", Woking Paper no. 5645, National Bureau of Economic Research.

Clements, M. T. (2004). "Low Quality as a Signal of High Quality", Working Paper, University of Texas at Austin.

Dewatripont, M., Jewitt., I., and Tirole, J. (1999). "The Economics of Career Concerns, Part I: Comparing Information Structures", *Review of Economic Studies*, **66**, 183-198.

Drazen, A. and Hubrich, S. (2003). "Mixed Signals in Defending the Exchange Rate: What do the Data Say?", CEPR Discussion Paper, no. 4050.

Edwards, R. (1976). "Individual Traits and Organizational Incentives: What Makes a Good Worker?", *Journal of Human Resourses*, **11**, 51-68.

Engers, M. (1987). "Signalling with Many Signals", *Econometrica*, **55**, 663-674.

Feltovich, N., Harbaugh R., and To, T. (2002). "Too Cool for School? Signaling and Countersignaling", *RAND Journal of Economics*, **33**, 630-649.

Grossman, S. (1981). "The Informational Role of Warranties and Private Disclosure About Product Quality", *Journal of Law and Economics*, **24**, 461-483.

Heckman, J., Stixrud, J. and Urzua, S. (2005) "The Effects of Cognitive and Noncognitive Abilities on Labor Market Outcomes and Social Behavior." Working Paper, University of Chicago.

Heckman, J. and Rubinstein, Y. (2001). "The Importance of Noncognitive Skills: Lessons from the GED Testing Program." *AEA Papers and Proceedings*, **91**, 145-9.

Holmstrom, B. (1999). "Managerial Incentive Problems: A Dynamic Perspective", *Review of Economic Studies*, **66**, 169-182.

Hvide, H. (2003). "Education and the Allocation of Talent", *Journal of Labor Economics*, **21**, 945-976.

Klein, R., Spady, R., and Weiss, A. (1991). "Factors Affecting The Output and Quit Propensities of Production Workers", *Review of Economic Studies*, **58**, 929-954.

Kholleppel, L. (1983). "Multidimensional 'Market Signaling" Discussion Paper, Universität Bonn.

O'Neal, B. (2002). "Nuclear Weapons and the Pursuit of Prestige", mimeograph, UCLA.

Okuno-Fujiwara, M., Postlewaite, A., and Suzumura, K. (1990). "Strategic Information Revelation", *Review of Economic Studies*, **57**, 25-47.

Orzach, R., Overgaard, P. B. and Tauman, Y. (2002). "Modest Advertising Signals Strength", *RAND Journal of Economics*, **33**, 340–358.

Pinker, S. (1999) *How the Mind Works* (Norton, New York, NY).

Quinzii, M. and Rochet, J. (1985). "Multidimensional Signaling", *Journal of Mathematical Economics*, **14**, 261-284.

Spence, M. (1973). *Market signalling: Information transfer in hiring and related processes* (Harvard

University Press, Cambridge, MA).

Sternberg, R. (1985). *Beyond IQ: A Triarchic Theory of Human Intelligence* (Cambridge University Press: Cambridge).

Riley, J. (1979). "Informational equilibrium", *Econometrica*, **47**, 331-359.

Tyler, J., Murnane, R., and Willet, J. (2000). "Estimating the Labor Market Signaling Value of the GED", *Quarterly Journal of Economics* **115**, 431-468.

Chapter 4

Multidimensional Incentive-Compatibility: The Multiplicatively Separable Case

4.1 Introduction

In many markets, some participants have information that is relevant to other participants. Frequently, information can be inferred from actions taken by the informed parties. The uninformed may move first and induce the informed to take such actions (screening) or the informed may move first and take actions in order to signal their information (signaling).

An important result in models with asymmetric information is the Revelation Principle, which states that any allocation process can be replicated by a mechanism in which participants are asked to reveal their private information. The Revelation Principle reduces a possibly complicated problem to an easy-to-state mathematical-programing problem, where each agent prefers to reveal his private information honestly (incentive-compatibility). However, the general analysis of such mathematical-programing problem is not straightforward.

Most of the literature assumes that an individual's private information consists of a one-dimensional type parameter and that the marginal utility of taking the action can be ordered

[0]This chapter is based on joint work with Aloisio Araujo and Humberto Moreira.

(single-crossing condition, SCC). Under this assumption, Mirrlees (1971), Spence (1974), and Rothschild and Stiglitz (1976) show that the solution of the programing problem is determined by a first-order and a monotonicity condition and the characterization of the set of incentive-compatible allocations becomes straightforward. McAfee and McMillan (1988) characterize incentive-compatible allocations in a multidimensional model under a single-crossing condition and Quinzii and Rochet (1985) characterize the separating equilibria of a multidimensional signaling model under a single-crossing condition.

This essay has two main purposes. First, it studies incentive-compatible allocations under a condition that is weaker than the SCC. This allows us to characterize the solution of multidimensional screening models as well as the equilibria of multidimensional signaling models where the SCC does not hold. Second, we determines the implications of multidimensional signaling and screening models when the SCC does not hold.

The characterization of optimal allocations in multidimensional screening models is complementary to Rochet (1987), McAfee and McMillan (1988), and Armstrong (1996). Our condition (discussed in Section 4.2) has three main advantages. First, it is easy to verify and is compatible with most specifications used in applied work (e.g. utility functions multiplicatively separable between the decision variable and types).[1] Second, it does not require assumptions to be made on endogenous variables and on the distribution of types. And third, it allows for utility functions that do not satisfy the SCC. The characterization of equilibria in multidimensional signaling models also provides necessary and sufficient conditions for signals to reveal all unobservable information (full-separability).[2]

Recently, the SCC has been criticized both on theoretical and empirical grounds. The first type of criticisms stress that, even though this property may be natural and intuitive in one-dimensional models, it cannot be extended in a sensible manner to multidimensional

[1] Armstrong (1996) assumes homogeneity of the utility function with respect to the type parameter, separability of the agent's indirect utility, and separability of the density function. McAfee and McMillan (1988) on the other hand assume a generalized single-crossing condition that requires the shadow-price indifference curves to be hyperplanes.

Matthews and Moore (1987) analyze a two-dimensional screening model where utility functions cross twice.

[2] Since Kholleppel's (1983) example of a model where no separating equilibrium existed, the existence of a fully-separating equilibrium became an important issue. Engers and Fernandez (1987) show that the SCC is sufficient for full-separability. The present article extends previous results on conditions required for full-separability by obtaining necessary and sufficient conditions.

settings. Models with discrete multidimensional types can be transformed into one-dimensional type models but the SCC is usually broken (see Rochet and Stole, 2003). Furthermore, utility functions that satisfy the analog of the SCC in multidimensional settings may fail to satisfy this condition in the presence of other signals (Araujo et al., 2007) or when there is correlation between the characteristics (Araujo and Moreira, 2003). Also, the presence of a moral hazard dimension may cause the SCC to be violated (Acemoglu, 1998).

In some applications, it is widely recognized that the assumption of one-dimensional types may be implausible. In the context of education, for example, Heckman (2005) argues that "abilities are multiple in nature", and that one-dimensional models cannot capture important phenomena (see Araujo et al., 2007, or Heckman et al., 2005).

Criticisms made on empirical grounds stress that several examples of interesting and intuitive phenomena have been proved to arise only when one drops this assumption. Bagwell and Bernheim (1996) investigate conditions under which consumers may be willing to pay higher prices for functionally equivalent goods as away to signal wealth ('Veblen effects'). Their main finding is that Veblen effects do not arise when the SCC holds but may arise when it fails.

Bernheim (1991) proves that firms may choose to distribute dividends even when they are taxed more highly than stock repurchases ('dividend puzzle') when the SCC fails. Bernheim and Severinov (2003) provide an explanation for the equal division of bequests based on a model where the SCC fails. Rotemberg (1988) shows that a tax on a signal may be Pareto-improving if the SCC fails to hold. Similarly, Bernheim and Redding (2001) show that taxing signals can be Pareto-improving in a model where the SCC does not hold. Bernheim (1994) studies a model of conformity in social interactions where the single-crossing condition fails.

The previous chapter of this thesis presented a model where high types choose to engage in a lower amount of signaling than intermediate types (see also Example 9). Araujo et al. (2004) show that the relation between dividend payments and earnings may be non-monotone when the SCC is not satisfied, and Smart (2000) and Araujo and Moreira (2003) show that the relation between risk and insurance coverage may not be monotonic.

Thus, one may question what the empirical content of the signaling and screening models is once the SCC is not assumed. In other words, which implications of these models are not a result of the particular specification of the cost of the activity. Similarly, it is important to understand

which results from the standard models generalize to multidimensional environments.[3]

We show that the only robust prediction of signaling is the monotonicity of transfers in costly actions. However, this prediction is shared by most alternative (symmetric information) models. Therefore, even when one employs a selection criterion, signaling can almost never be rejected. Another negative result concerns the identifiably of signaling models. It is shown that, for each signaling model, there exists a large class of signaling models with the same observable implications. Hence, it is impossible to determine which is correct among a large class of models.

The characterization of solutions of multidimensional screening allows us to identify a new necessary and sufficient condition. Under homogeneity of the distribution of types or when types are one-dimensional, this condition implies that the principal's profit must grow with respect to types at a higher rate under asymmetric information than under symmetric information.

Overall, our results imply that special attention must be devoted to the specific characteristics of the market being studied. Only with precise knowledge about the cost of engaging in the activity, the relevant number of dimensions, and the technology, one is able to obtain testable predictions of the model.

This chapter is also related to the literature on monotone comparative statics, which studies necessary and sufficient conditions for solutions to maximization programs to be monotone. One important result in this literature is that the SCC is sufficient for the solution to be monotone (c.f., Milgrom and Shannon, 1994). Furthermore, if the choice set is sufficiently rich, this condition is also necessary. In the present paper, we show that the cross-partial derivative of the cost function determines not only whether the set of incentive-compatible allocations is increasing or decreasing but the whole shape of the solution.

The rest of the chapter is organized as follows. Section 4.2 characterizes the set of incentive-compatible allocations and studies the restrictions imposed by incentive-compatibility. Section 4.3 presents the screening model and characterizes the optimal solution. Then, we study the additional implications imposed by the screening model. Section 4.4 considers the signaling

[3]In a model of insurance under asymmetric information, Chiappori et al. (2006) argue that coverage and risk are positively correlated even when the SCC is not imposed or when types are multidimensional. However, although we also obtain a monotonicity condition, our results are not directly applicable to insurance since we follow the standard mechanism design model and assume quasi-linear utilities.

model. Then, Section 4.5 concludes.

4.2 Characterization of Incentive Compatibility

The economy consists of informed agents and an uninformed principal. Each agent is characterized by a multidimensional parameter ('type') $\theta \in \Theta$, where $\Theta \subset \mathbb{R}_+^n$ is a compact and convex set with non-empty interior. Types are distributed according to a continuous density $p : \Theta \to \mathbb{R}_{++}$. Agents may engage in a costly activity $y \in \mathbb{R}_+$. Principals have a continuously differentiable valuation function $f : \Theta \times \mathbb{R}_+ \to \mathbb{R}_+$.

The cost of the activity y is given by a C^3 function $c : \Theta \times \mathbb{R}_+ \to \mathbb{R}_+$, where

Assumption 0 $c_y(\theta, y) > 0$ and $c_{\theta_i}(\theta, y) < 0$, for all $\theta \in \Theta$, $y \in \mathbb{R}_+$, and $i = 1, ..., n$.

The first inequality states that y is costly while the second states that higher types have a lower cost of engaging in the activity y.

The standard single-crossing condition (SCC) requires $c_{\theta_i y}$ not to change signs so that having a higher type has a monotonic effect on the marginal cost of engaging in y. The following assumption generalizes the SCC in the sense that it allows $c_{\theta_i y}$ to change sign.

Assumption 1 There exist functions $\xi : \mathbb{R}_+ \to \mathbb{R}$, $\psi : \Theta \to \mathbb{R}$, and $\varphi : \Theta \to \mathbb{R}$ such that

$$c(\theta, y) = \xi(y) + y \times \psi(\theta) + \varphi(\theta).$$

Assumption 1 implies that $c_{\theta_i yy}(\theta, y) = 0$, for all $i = 1, ..., n$. It is satisfied, for example, when costs are quadratic in y :

$$\hat{c}(\theta, y) = J \times y^2 + y \times \psi(\theta) + \varphi(\theta),$$

where J is a real number. This representation includes the standard case of costs that are linear in y ($J = \varphi(\theta) = 0$ for all θ).

The SCC holds in each dimension if either $c_{\theta_i y}(\theta, y) > 0$ or $c_{\theta_i y}(\theta, y) < 0$ for all y, θ_i, and i. Assumption 1 states that $c_{\theta_i y}(\theta, y)$ is not a function of y. Hence, if we fix the $n - 1$ other dimensions and consider a graph of θ_i and y, $c_{\theta_i y}$ is constant along any vertical line. In

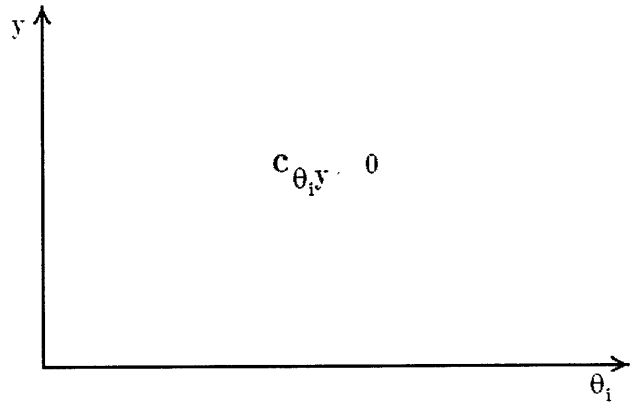

Figure 4-1: Cost function that satisfies SCC in each dimension.

particular, the intervals where $c_{\theta_i y} > 0$ and where $c_{\theta_i y} < 0$ are separated by vertical lines (see Figures 4-1 and 4-2).

In a context of education as signal, for example, the agents are workers, the principals are employers, the activity y is the amount of schooling, and the price w is the wage paid to workers. In an industrial organization context, the agents are firms, principals are potential buyers, and the activity may be the amount spent on advertisement or the duration of warranties. The following example considers the model from Chapter 3.

Example 9 *Consider a labor market model. Risk neutral workers engage in a costly schooling activity* $y \in \mathbb{R}_+$ *and obtain wages* $w \in \mathbb{R}_+$. *Their ability is captured by a two-dimensional vector* $\theta \equiv (\theta_1, \theta_2) \in \mathbb{R}^2$ *and the cost of schooling is*[4]

$$c(\theta, y) = \frac{y}{\theta_1 \theta_2}.$$

Employers do not observe ability but observe the amount of schooling y. *They can also interview workers, which gives them an additional measure of each worker's skills. The interview*

[4] θ_1 *and* θ_2 *can be interpreted as cognitive and non-cognitive skills, for example.*

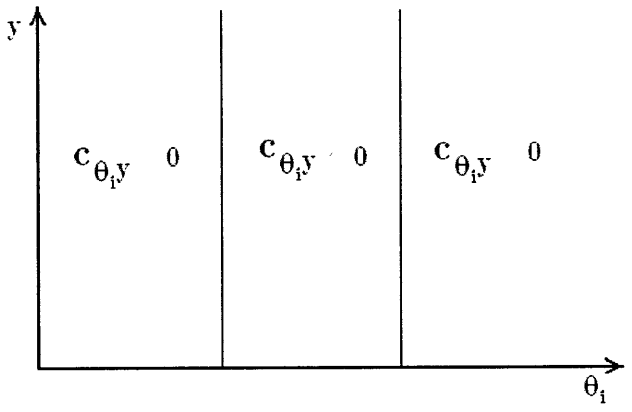

Figure 4-2: Cost function that satisfies Assumption 1.

technology is represented by the following function:

$$g(\theta, y) = \alpha\theta_1 + \theta_2 + \beta y.$$

A type-θ worker produces a good that is worth $f(y, \theta)$, where $f_{\theta_i} > 0$, $f_y \geq 0$, $f_{yy} \leq 0$, and $f_{y\theta_i} \geq 0$. For a fixed $g(\theta, y) = \bar{g}$, we can write the cost of schooling as

$$\hat{c}(\theta_1, y, \bar{g}) \equiv \frac{y}{\theta_1 (\bar{g} - \alpha\theta_1 - \beta y)}.$$

Note that the single-crossing condition does not hold since $\hat{c}_{\theta_1 y} \{>_<\} 0 \iff \theta_1 \{>_<\} \frac{\bar{g} - \beta y}{2\alpha}$.

The single-crossing condition means that exchanging one unit of ability θ_1 for α units of θ_2 would always be either desirable or undesirable in terms of reducing the cost of schooling. In the specification above, because the abilities are imperfect substitutes, this exchange is desirable for high levels of θ_1 and undesirable for low levels of θ_1. Therefore, the substitutability between skills breaks down the single-crossing condition. Furthermore, Assumption 1 is satisfied when $\beta = 0$.

The next example describes a model of warranties:

Example 10 *Consider a model of warranties and uncertain product quality. Product quality is determined by a two-dimensional vector of characteristics $\theta \equiv (\theta_1, \theta_2) \in \mathbb{R}^2$. For concreteness,*

interpret θ_1 as the reliability and θ_2 as the complexity of the good. Consumers observe a measure of product quality given by the function $g(\theta) = \alpha\theta_1 + \theta_2$.

Producers may offer a warranty that repairs any defect that may occur until time y. Let $w(y, g)$ denote the price charged conditional on $g(\theta) = g$ when warranty y is provided.

Denote by $f(\theta_1, \theta_2, y)$ the expected value of the good to consumers given warranty y and $c(\theta_1, \theta_2, y)$ denote the expected cost of producing the good and providing warranty y. We assume that reliability reduces the cost of providing warranty whereas complexity increases the cost of providing warranty:

$$c_{\theta_1 y} < 0, \ c_{\theta_2 y} > 0.$$

As in Example 9, we can rewrite the expected cost of producing the good conditional on $\bar{g}$ as $\hat{c}(\theta_1, y, \bar{g}) \equiv c(\theta_1, \alpha\theta_1 - \bar{g}, y)$. Note that $\hat{c}_{\theta_1 y} = c_{\theta_1 y} + \alpha c_{\theta_2 y}$ may switch signs because the first term is negative while the second term is positive. Therefore, the fact that reliability decreases the cost of providing warranty whereas complexity increases this cost implies that single-crossing condition may not hold. In particular, if we assume linear costs,

$$c(\theta_1, \theta_2, y) = A \times \theta_1 \times (K - \theta_2) \times y,$$

where A and K are positive real numbers, it follows that $\hat{c}_{\theta_1 y} \left\{ \begin{matrix} > \\ < \end{matrix} \right\} 0 \iff \theta_1 \left\{ \begin{matrix} < \\ > \end{matrix} \right\} \frac{K + \bar{g}}{2\alpha}$ and Assumption 1 is satisfied.

The following example presents a multidimensional generalization of the standard nonlinear pricing model:

Example 11 *A monopolist sells a good in different sizes (or qualities) $Q \geq 0$. Consumers have private information about their tastes, captured by a vector of types $\theta \in \Theta$. The consumer's gross surplus from the good is $V(\theta, Q)$, where $\frac{\partial V}{\partial \theta_i} > 0$ and $\frac{\partial V}{\partial Q} > 0$. A purchase of size Q is sold at price $P(Q)$. Therefore, the consumer's utility from purchasing the good is $V(\theta, Q) - P(Q)$.*

The cost of production is $FC + MC \times Q$, where $FC \geq 0$ and $MC > 0$ denote the firm's fixed and marginal costs. The firm's profit from selling the good is $P(Q) - MC \times Q - FC$.

Let $y = \bar{Q} - Q$ for $\bar{Q}$ large enough, let $w(y) = -P\left(\bar{Q} - y\right)$, and define the function

$c : \Theta \times \mathbb{R}_+ \to \mathbb{R}_+$ *by*

$$c(\theta, y) = -V\left(\theta, \bar{Q} - y\right).$$

The consumer's utility can be written as $w(y) - c(\theta, y)$. *Furthermore,* $c(\theta, y)$ *satisfies Assumption 0.*

Let $f(y) = -MC \times \left(\bar{Q} - y\right) - FC$. *The firm's profit can be written as* $f(y) - w(y)$. *Hence, this model is a special case of the basic framework. Moreover, Assumption 1 is satisfied if and only if there exist functions* $\xi : \mathbb{R}_+ \to \mathbb{R}$, $\psi : \Theta \to \mathbb{R}$, *and* $\varphi : \Theta \to \mathbb{R}$ *such that*

$$V(\theta, Q) = \xi(Q) + Q \times \psi(\theta) + \varphi(\theta).$$

The next example describes of the random participation model of Rochet and Stole (2002):

Example 12 *Consider the same setting as in Example 11. Take* $\theta = (\theta_1, \theta_2)$ *and take* $\xi(Q) = 0$, $\psi(\theta) = \theta_1$, $\varphi(\theta) = -\theta_2$. *The parameter* θ_1 *denotes the consumer's taste for quality* Q, *and* θ_2 *indexes the consumer's opportunity cost.*

A mechanism $(y(\theta), w(y))$ is incentive-compatible if it satisfies the following incentive-compatibility constraint:

$$y(\theta) \in \arg\max_{\widehat{y}} w(\widehat{y}) - c(\theta, \widehat{y}) \tag{IC}$$

for all $\theta \in \Theta$. Equivalently, we say that $w(y)$ implements $y(\cdot)$. We say that a profile of activities $y(\cdot)$ is implementable if there exists a function $w(y)$ that implements it.

Before proceeding, we need to introduce some notation. We say that a type θ is regular for $y(\cdot)$ if the differential $Dy(\theta)$ has full rank; otherwise θ is critical. Finally, $\overline{y}$ is a critical value of y if there exists a critical type θ such that $y(\theta) = \overline{y}$, otherwise $\overline{y}$ is called a regular value of y. We refer to a C^1 function with measure zero of critical values as a *regular function.* In what follows we will only consider mechanisms $(y(\theta), w(y))$ such that y and w are regular.[5,6]

[5] Apart from the differentiability assumption, the only restrictions that this condition imposes to the incentive compatible profile is that there are no discontinuities in $y(\cdot)$ and its derivative. The results could be generalized with an adaptation in the proof of Theorem 1: the price profile would have kinks (which would correspond to the discontinuities of the derivative of y) and disconnected domain (which would correspond to the discontinuities in y).

[6] By Sard's theorem, this condition is automatically satisfied if $y \in C^n$, $n > 1$.

Which mechanisms $(y(\theta), w(y))$ satisfy incentive-compatibility? In the remainder of this section, we first analyze necessity and then sufficiency. The following lemma gives the usual first- and second-order conditions of the incentive compatibility constraint.

Lemma 10 *Let $(y(\cdot), w(\cdot))$ be an incentive-compatible mechanism. Then,*

$$w'(y(\theta)) = c_y(\theta, y(\theta)), \tag{4.1}$$

$$c_{\theta_i y}(\theta, y(\theta))\, y_{\theta_i}(\theta) \leq 0, \tag{4.2}$$

for all $i = 1, ..., n$ and $\theta \in \Theta$.

Proof. To simplify the proof, let us assume that $y(\cdot)$ is C^2. If $(y(\cdot), w(\cdot))$ is incentive-compatible, it must satisfy:

$$\theta \in \arg\max_{\hat{\theta}} w\left(y\left(\hat{\theta}\right)\right) - c\left(\theta, y\left(\hat{\theta}\right)\right),$$

whose local first- and second-order conditions are

$$\begin{aligned} w'(y(\theta))\, y_{\theta_i}(\theta) - c_y(\theta, y(\theta))\, y_{\theta_i}(\theta) &= 0, \\ w''(y(\theta))\,[y_{\theta_i}(\theta)]^2 + w'(y(\theta))\, y_{\theta_i\theta_i}(\theta) - c_{yy}(\theta, y(\theta))\,[y_{\theta_i}(\theta)]^2 - c_y(\theta, y(\theta))\, y_{\theta_i\theta_i}(\theta) &\leq 0. \end{aligned}$$

Equation (4.1) follows from the first-order condition. Differentiating the first-order condition (which must hold for every θ) yields

$$\begin{aligned} & w''(y(\theta))\,[y_{\theta_i}(\theta)]^2 + w'(y(\theta))\, y_{\theta,\theta_i}(\theta) - c_{yy}(\theta, y(\theta))\,[y_{\theta_i}(\theta)]^2 - c_y(\theta, y(\theta))\, y_{\theta_i\theta_i}(\theta) \\ = \; & c_{\theta_i y}(\theta, y(\theta))\, y_{\theta_i}(\theta). \end{aligned}$$

Substituting in the second order condition, we obtain $c_{\theta_i y}(\theta, y(\theta))\, y_{\theta_i}(\theta) \leq 0$. ■

When the SCC holds, equation (4.2) reduces to a monotonicity condition. When it does not hold, equation (4.2) implies that incentive-compatible mechanisms may not be monotonic: y has to be increasing in the region where $c_{\theta,y} < 0$ and decreasing in the region where $c_{\theta,y} > 0$.

In one-dimensional models where mechanisms are monotone, if two types pool in the activity

y, all intermediate types must also be pooled with them. As a consequence, pooling sets must be intervals. However, when mechanisms are not monotone, a disconnected set of types may be pooled. Araujo and Moreira (2004) have shown that a necessary condition for incentive-compatibility in this case is the so-called marginal utility identity. This condition implies that if two disconnected types are pooling in an activity y, they should have the same marginal cost. The following lemma establishes this result in our context:

Lemma 11 *If two individuals with regular types θ and $\tilde{\theta}$ choose the same signal, then their marginal cost must be the same:*

$$y(\theta) = y\left(\tilde{\theta}\right) \Rightarrow c_y(\theta, y(\theta)) = c_y\left(\tilde{\theta}, y(\theta)\right). \tag{4.3}$$

Proof. Suppose that $(w(y(\theta)), y(\theta))$ is incentive-compatible. Then, (4.1) must hold. Therefore, if θ and $\widetilde{\theta}$ are regular types such that $y(\theta) = y\left(\tilde{\theta}\right)$, equation (4.1) implies that

$$c_y(\theta, y(\theta)) = c_y\left(\tilde{\theta}, y\left(\tilde{\theta}\right)\right).$$

■

We have shown that the local conditions (4.1), (4.2) and the global condition (4.3) are necessary for incentive-compatibility even when Assumption 1 does not hold. The following example shows that they may not be sufficient when Assumption 1 does not hold.

Example 13 *Suppose that the cost of activity y is*

$$c(\theta, y) = \frac{y}{2|\theta - 1|} + \frac{1}{8}\left[y - (\theta - 1)^2\right]^2 - \frac{5\theta}{3}\left[y - (\theta - 1)^2\right]^3,$$

so that Assumption 1 is not satisfied. Consider the following mechanism:

$$\begin{aligned} y(\theta) &= (\theta - 1)^2, \\ w(y) &= \sqrt{y}, \end{aligned}$$

where $\Theta = [1.1,\ 2]$.

Note that conditions (4.1), (4.2), and (4.3) are satisfied. Under the proposed mechanism, type $\theta = \frac{3}{2}$ *chooses* $y\left(\frac{3}{2}\right) = \frac{1}{4}$ *and obtains utility* $w\left(\frac{1}{4}\right) - c\left(\frac{3}{2}, \frac{1}{4}\right) = \frac{1}{4}$. *However, by choosing* $y = 1$, *he obtains* $w(1) - c\left(\frac{3}{2}, 1\right) = \frac{63}{64} > \frac{1}{4}$, *which is a profitable deviation. Therefore, (4.1), (4.2), and (4.3) may not prevent non-local deviations when Assumption 1 does not hold.*

The proposition below establishes that, under Assumption 1, the necessary conditions are sufficient as well.

Proposition 26 *Suppose Assumption 1 holds. A mechanism* $(y(\cdot), w(\cdot))$ *is incentive-compatible if and only if the first- and second-order conditions (4.1) and (4.2), and the pooling condition (4.3) are satisfied.*

Proof. ($\Rightarrow$) Follows from Lemmata 10 and 11.

($\Leftarrow$) Let us define the wage schedule $w \in C^1$. By (4.3) we can define the derivative of w by $w'(y) = c_y(\theta, y)$ for $y = y(\theta)$ and all regular types $\theta \in \Theta$. Since the set of critical values has zero measure, we can extend continuously the definition of $w'(y)$ for critical values. Then, define the following wage schedule for $y \in y(\Theta)$:[7]

$$w(y) = \int_{y^{\min}}^{y} w'(x)dx + w^{\min},$$

where $w^{\min} \geq \max_{\theta \in \Theta} c(\theta, 0)$ (which ensures that all types will participate) and $y^{\min} = \min_{\theta \in \Theta} y(\Theta)$.

Let θ be first a regular type (for critical θ the argument is made by continuity). The first-order condition of the previous problem is $w'(y(\theta)) = c_y(\theta, y(\theta))$, which holds by the definition of $w(\cdot)$.

Given a regular $\widehat{y} = y(\widehat{\theta})$, there exists $i \in \{1, ..., n\}$ such that $y_{\theta_i}(\widehat{\theta}) \neq 0$. By the implicit function theorem there exists a function φ such that, locally, $\widehat{\theta}_i = \varphi(\widehat{y}, \widehat{\theta}_{-i})$, where $\widehat{\theta}_{-i}$ is the vector $n-1$ dimensional $\widehat{\theta}$ but coordinate i. Taking the derivative with respect to $\widehat{y}$ in the equality $w'(\widehat{y}) = c_y(\varphi(\widehat{y}, \widehat{\theta}_{-i}), \widehat{\theta}_{-i}, \widehat{y})$ we get

$$w''(\widehat{y}) = c_{yy}(\widehat{y}, \widehat{\theta}) + c_{\theta_i y}\left(\varphi\left(\hat{y}, \widehat{\theta}_{-i}\right), \widehat{\theta}_{-i}, \hat{y}\right) \varphi_y\left(\hat{y}, \widehat{\theta}_{-i}\right).$$

[7] For $y \notin y(\Theta)$, we extend the wage function linearly such that the derivative is continuous.

Therefore, the second derivative of the agent's utility at $\widehat{y}$ is

$$w''(\widehat{y}) - c_{yy}(\theta, \widehat{y}) = c_{yy}(\widehat{\theta}, \widehat{y}) - c_{yy}(\theta, \widehat{y}) + c_{\theta_i y}\left(\varphi\left(\hat{y}, \widehat{\theta}_{-i}\right), \widehat{\theta}_{-i}, \hat{y}\right)\varphi_y\left(\hat{y}, \widehat{\theta}_{-i}\right).$$

However, Assumption 1 implies that $c_{yy}(\widehat{\theta}, \widehat{y}) - c_{yy}(\theta, \widehat{y}) = 0$ because $c_{\theta_i yy} = 0$ for all i. Then, the sign of the second derivative is given by (4.2), i.e., it is negative. Again, for critical $\widehat{y}$ we can use continuity. This implies that $y(\theta)$ is the maximum on $y(\Theta)$. This concludes the proof.

■

Next, we use the characterization from Proposition 26 to study implications of incentive-compatibility when the SCC is not imposed. Our assumption that the activity y is costly implies that transfers must be strictly increasing in y. Therefore, $w'(y) > 0$ is a necessary condition for incentive-compatibility. The theorem below states that it is also a sufficient condition. More specifically, given any mechanism $(y(\theta), w(y))$ we can find a cost function satisfying Assumptions 0 and 1 for which such schedule is incentive-compatible.

Theorem 1 *Let $y(\cdot)$ be a regular function and let $w(\cdot)$ be a positive C^2 function. There exists a C^1 cost function satisfying Assumption 1 for which $(y(\cdot), w(\cdot))$ is incentive-compatible if and only if $w(\cdot)$ is strictly increasing. Moreover, if $w(\cdot)$ is concave, such cost function can be chosen such that Assumption 0 is also satisfied.*

Proof. ($\Rightarrow$) Follows from revealed preference.

($\Leftarrow$) Let us define the following C^1 function:

$$c(\theta, y) = A(\theta) + w'(y(\theta))y + \frac{K}{2}(y - y(\theta))^2,$$

where $K > 0$ is a constant such that $w''(y(\theta)) < K$ and $A(\theta)$ is such that $c(\theta, y) > 0$ and $c_{\theta_i}(\theta, y) < 0$. Such function $A(\theta)$ exists.

The function c is C^1 because $y(\Theta)$ is a compact set and $w(\cdot)$ is C^2. Moreover, the marginal cost $c_y(\theta, y) = w'(y(\theta)) + K(y - y(\theta))$ is always positive along $y = y(\theta)$. A sufficient condition for the existence of a K such that the marginal cost is positive for all $y \geq 0$ and $w''(y(\theta)) < K$ is $w''(y(\theta)) < \frac{w'(y(\theta))}{y(\theta)}$ for all θ. This condition is obviously satisfied when the wage function is concave since $w' > 0$ and Θ is compact.

Note that c satisfies Assumption 1. We claim that the pair $(y(\cdot), w(\cdot))$ is incentive-compatible. First, (4.1) is trivial and (4.2) holds because

$$c_{\theta,y}(\theta, y(\theta))y_{\theta_i}(\theta) = \left[w''(y(\theta)) - K\right]\left[y_{\theta_i}(\theta)\right]^2 \leq 0.$$

Furthermore, if θ and $\widetilde{\theta}$ are regular values of $y(\cdot)$, then

$$\begin{aligned} c_y(\theta, y) = c_y(\widetilde{\theta}, y) &\iff w'(y(\theta)) - Ky(\theta) = w'(y(\widetilde{\theta})) - Ky(\widetilde{\theta}) \\ &\iff y = y(\theta) = y(\widetilde{\theta}), \end{aligned}$$

since $w'(y) - Ky$ is decreasing on y (by the assumption on K), i.e., (4.3) holds. Using Proposition 26, we conclude the proof. ■

The theorem above implies that, apart from the monotonicity of the transfer function, incentive-compatibility by itself does not lead to any additional restrictions on the space of incentive-compatible mechanisms. In the next sections, we characterize the solutions of screening and signaling models and analyze which additional restrictions arise.

4.3 The Screening Model

In this section, we embed the structure of the previous section into a screening model. There is one uninformed principal who makes a take-it-or-leave-it offer to an informed agent. The agent's private information is characterized by the parameter θ, which is distributed according to the density p.

The revelation principle allows us to restrict the space of contracts to direct mechanisms that satisfy the incentive-compatibility constraint (IC). Each type has an outside option that gives a constant reservation utility normalized to zero.[8] Therefore, the principal faces the following participation constraint:

$$w(y(\theta)) - c(\theta, y(\theta)) \geq 0 \qquad \forall \theta \in \Theta. \tag{IR}$$

The following definition states the principal's problem:

[8] One could also allow for type-dependent reservation utilities. For the purposes of our results on the additional restrictions imposed by screening, we can always ajust the cost function to avoid countervailing incentives.

Definition 8 *The principal's program is*

$$\max_{(y(\theta),w(y))} E[f(\theta, y(\theta)) - w(y(\theta))] \tag{4.4}$$

$$s.t.\ (IC)\ and\ (IR).$$

Proposition 26 states that, under Assumption 1, the space of incentive-compatible mechanisms is characterized by conditions (4.1), (4.2) and (4.3). Define the agent's informational rent given an incentive-compatible mechanism $(y(\theta), w(y))$ as

$$r(\theta) \equiv w(y(\theta)) - c(\theta, y(\theta)). \tag{4.5}$$

Applying the Envelope Theorem (see Milgrom and Segal, 2002), we obtain the following condition:

$$\nabla r(\theta) = -\nabla_\theta c(\theta, y(\theta)), \tag{4.6}$$

where ∇ is the gradient operator. Note that condition (4.6) is equivalent to (4.1).

In order to solve the principal's program, we follow the standard approach of considering first a relaxed program which ignores some of the constraints. Then, we state conditions that ensure that the ignored constraints do not bind so that the solution of the relaxed program is the same as the solution of the principal's program.

The relaxed program is defined as the maximization of the principal's profit subject to the agent's first-order condition and the participation constraint:

$$\max_{(y(\theta),w(y))} E[f(\theta, y(\theta)) - w(y(\theta))]$$

$$s.t.\ (4.1)\ \text{and (IR)}.$$

Note that the program above does not take into account the local second-order conditions (4.2) and the global conditions (4.3). For clarity, it is convenient to analyze the one-dimensional and the multidimensional cases separately.

4.3.1 The one-dimensional case

This subsection characterizes the solutions of screening models when the parameter of private information θ is one-dimensional and considers their empirical implications. We obtain a new necessary and sufficient condition for a mechanism to be a solution to a screening model. The new condition, which does not depend on the SCC, states that the principal's profit must increase in the agent's type at a higher rate under asymmetric information than in the symmetric information case. Equivalently, the condition identifies the regions where the principal's profit is increasing and decreasing in the activity y chosen by the agent.

From our assumption about the type space, it can be written as $\Theta = \left[\underline{\theta}, \overline{\theta}\right] \subset \mathbb{R}$. Note that Assumption 0 implies that the informational rent is increasing in the agent's type. Therefore, the participation constraint (IR) is satisfied if and only if $r\left(\underline{\theta}\right) \geq 0$. Furthermore, in the solution to Program (4.4) it must be the case that $r(\underline{\theta}) = 0$.

Integrating (4.6), the agent's informational rent becomes

$$r(\theta) = r(\underline{\theta}) - \int_{\underline{\theta}}^{\theta} c_{\theta}(\widetilde{\theta}, y(\widetilde{\theta}))d\widetilde{\theta}.$$

Applying integration by parts, we obtain

$$E[w(y(\theta))] = r(\underline{\theta}) + E\left[c(\theta, y(\theta)) - \frac{1 - P(\theta)}{p(\theta)} c_{\theta}(\theta, y(\theta))\right],$$

for every incentive-compatible mechanism $(y(\theta), w(y))$.

Substituting into the objective function, the relaxed program becomes:

$$\max_{(y(\theta), w(y))} E\left[f\left(\theta, y\left(\theta\right)\right) - c\left(\theta, y\left(\theta\right)\right) + \frac{1 - P(\theta)}{p(\theta)} c_{\theta}\left(\theta, y\left(\theta\right)\right)\right].$$

The pointwise first-order condition of the relaxed program is equivalent to

$$f_y\left(\theta, y\left(\theta\right)\right) - c_y\left(\theta, y\left(\theta\right)\right) + \frac{1 - P(\theta)}{p(\theta)} c_{\theta y}\left(\theta, y\left(\theta\right)\right) = 0. \tag{4.7}$$

This condition depicts the usual trade-off between rent extraction and distortion that the principal faces. Instead of equating the marginal valuation to the marginal cost, the principal sets

it equal to the marginal cost plus the marginal cost of information rents.

For simplicity, suppose that w is C^2. Taking the total derivative of equation (4.1) with respect to θ gives:

$$w''(y(\theta))y'(\theta) = c_{\theta y}(\theta, y(\theta)) + c_{yy}(\theta, y(\theta))y'(\theta).$$

Thus, (4.7) can be written as

$$f_y(\theta, y(\theta)) - w'(y(\theta)) + \frac{1 - P(\theta)}{p(\theta)}[w''(y(\theta)) - c_{yy}(\theta, y(\theta))]y'(\theta) = 0. \tag{4.8}$$

Note that the (necessary) local second-order condition of Program (IC) is $w''(y(\theta)) - c_{yy}(\theta, y(\theta)) \leq 0$. Substituting from equation (4.8), it follows that

$$[f_y(\theta, y(\theta)) - w'(y(\theta))]y'(\theta) \geq 0, \text{ for all } \theta \in \Theta, \tag{4.9}$$

since the hazard rate is always positive.

Condition (4.9) has a natural interpretation in terms of the principal's profit: $\pi(\theta) \equiv f(\theta, y(\theta)) - w(y(\theta))$. Under symmetric information, we would have $\pi'(\theta) = f_\theta(\theta, y(\theta))$. Differentiating the profit function, we obtain:

$$\pi'(\theta) = f_\theta(\theta, y(\theta)) + [f_y(\theta, y(\theta)) - w'(y(\theta))]\, y'(\theta) \geq f_\theta(\theta, y(\theta)),$$

where the inequality uses equation (4.9). Therefore, under asymmetric information, the principal's profit increases at a greater rate than the increase in productivity. Note that this result is quite general in that it does not depend on any assumptions on the cost function except for it being increasing in y and decreasing in θ and the assumption that types are one-dimensional. The next subsection shows that this result can be somewhat generalized to multidimensional types.

Condition (4.9) can also be interpreted as determining the effect of activity y on the principal's profit. From condition (4.2), it can be written as:

$$f_y(\theta, y(\theta)) \left\{ \begin{array}{c} \geq \\ \leq \end{array} \right\} w'(y(\theta)) \iff c_{\theta y}(\theta, y(\theta)) \left\{ \begin{array}{c} \leq \\ \geq \end{array} \right\} 0. \tag{4.10}$$

Thus, the profit must be increasing in y in the region where the type decreases the marginal cost of the activity (controlling for the type θ).[9] Conversely, the profit must be decreasing when types increase the marginal cost of the activity y. In the standard case where $c_{\theta y} < 0$ (the single-crossing condition is satisfied), it implies that the principal's profit is increasing in the activity y controlling for the agent's type.

The previous argument shows the necessity of condition (4.9). It turns out that this condition is also sufficient to rationalize any incentive-compatible mechanism as the solution of the principal's program when Assumption 1 is satisfied:

Theorem 2 *Suppose $\Theta \subset \mathbb{R}$. Let $y(\cdot)$ be a regular function and let $w(\cdot)$ be a positive C^2 function. There exists a C^1 cost function satisfying Assumption 1 and a distribution of types p for which $(y(\cdot), w(\cdot))$ is the optimal mechanism if and only if $w(\cdot)$ is strictly increasing and condition (4.9) is satisfied.*

Proof. ($\Rightarrow$) Follows from the preceding argument and Theorem 1.

($\Leftarrow$) Let $(y(\theta), w(y))$ be a mechanism satisfying the conditions of the theorem. Take the cost function of the proof of Theorem 1. Then, the mechanism satisfies incentive-compatibility and equation (4.7) becomes:

$$\frac{f_y(\theta, y(\theta)) - w'(y(\theta))}{[w''(y(\theta)) - K]y'(\theta)} + \frac{1 - P(\theta)}{p(\theta)} = 0. \tag{4.11}$$

Hence, by (4.9) we can then define the following function:

$$P(\theta) = 1 - A \times \exp\left[\int_{\theta}^{\overline{\theta}} \frac{K - w''(y(t))}{f_y(t, y(t)) - w'(y(t))} y'(t) dt\right],$$

where A is chosen such that $P(\underline{\theta}) = 0$. It is easy to see that $P(\cdot)$ is a cumulative distribution function which satisfies (4.11). Note that for such economy the second-order condition of the relaxed functional holds if and only if

$$f_{yy}(\theta, y(\theta)) - K \leq 0.$$

[9] By 'controlling for the type', we mean the effect on the profit if an agent chose a different amount of activity y. Of course, in equilibrium we only observe one action for each type: $y = y(\theta)$.

Therefore, choosing K with this property we conclude that $(y(\theta), w(y))$ is the solution of the principal's program.[10] ■

4.3.2 The multidimensional case

As in McAfee and McMillan (1988) and Rochet (1987), this subsection obtains necessary and sufficient conditions for implementability and optimality in a multidimensional screening model. The implementability result follows straight from Proposition 26, which characterizes the set of incentive-compatible mechanisms under Assumption 1. In this subsection, we use this result to characterize the optimal mechanism.

In general, the first difficulty in multidimensional screening is to deal with the integration of equation (4.6). In order to deal with this, we will follow the approach proposed by Armstrong (1996).

Assume that there exists a $\underline{\theta} \in \Theta$ such that $\underline{\theta} \leq \theta$ for all $\theta \in \Theta$. With no loss of generality, take $\underline{\theta} = 0$.[11] Under Assumption 0, the informational rent (4.5) is increasing in the agent's type. Thus, the participation constraint is satisfied if and only if $r(0) \geq 0$ so that, in the solution to Program (4.4), we must have $r(0) = 0$.

Consider the expected value of the agent's informational rent (4.5):

$$R = \int_{\Theta} r(\theta)p(\theta)d\theta.$$

Define the function $v : [0,1] \to \mathbb{R}$ by $v(t) = \int_{\Theta} r(t\theta)p(\theta)d\theta$. Then, it follows that $v'(t) = \int_{\Theta} \theta \cdot \nabla r(t\theta)p(\theta)d\theta$, where $\cdot$ denotes the inner product. Because $r(0) = 0$, we have

$$v(0) = 0 \text{ and } v(1) = R.$$

Since $v(1) - v(0) = \int_{\Theta} v'(t)dt$, it follows that

$$R = \int_0^1 \left(\int_{\Theta} \theta \cdot \nabla r(t\theta)p(\theta)d\theta \right) dt. \tag{4.12}$$

[10] We can choose a cost function such that the marginal cost is always positive whenever $f_{yy}(\theta, y(\theta)) < \frac{w'(y(\theta))}{y(\theta)}$. A sufficient condition is that w is more concave than $f(\theta, \cdot)$ for all θ, i.e., $f_{yy}(\theta, y) < w''(y)$ for all $y \geq 0$ (see the proof of Theorem 1).

[11] This can be obtained by redefining the types as $\tilde{\theta} = \theta - \underline{\theta}$.

The envelope condition (4.6) implies that $R = -\int_0^1 \left(\int_\Theta \theta \cdot \nabla_\theta c(t\theta, y(t\theta))p(\theta)d\theta\right) dt$. For each $t > 0$, we apply the change of variables $\eta = t\theta$, which takes Θ into itself. Under this transformation, the term $\theta_i c_{\theta_i}(t\theta, y(t\theta))$ becomes

$$\frac{1}{t}\eta_i c_{\theta_i}(\eta, y(\eta)).$$

Then, the integral with respect to θ in equation (4.12) can be transformed according to

$$\int_\Theta \theta \cdot \nabla_\theta c(t\theta, y(t\theta))p(\theta)d\theta = t^{-n-1}\int_\Theta \eta \cdot \nabla_\theta c(\eta, y(\eta))p\left(\frac{\eta}{t}\right) d\eta,$$

so that the expected informational rent becomes

$$R = -\int_\Theta \eta \cdot \nabla_\theta c(\eta, y(\eta)) \left(\int_0^1 t^{-n-1} p\left(\frac{\eta}{t}\right) dt\right) d\eta.$$

Finally, letting $\tau = 1/t$ and defining $q(\eta) = \int_1^\infty \tau^{n-1}p(\tau\eta)d\tau$, we obtain

$$R = -\int_\Theta \eta \cdot \nabla_\theta c(\eta, y(\eta))q(\eta)d\eta.$$

Substituting R back into the principal's objective function, we obtain the following relaxed maximization problem

$$\max_{y(\cdot)} E\left[f(\theta, y(\theta)) - c(\theta, y(\theta)) + \theta \cdot \nabla_\theta c\left(\theta, y\left(\theta\right)\right)\frac{q(\theta)}{p(\theta)}\right], \tag{4.13}$$

where the expectation operator is taken with respect to the probability measure defined by the density $p(\theta)$.

The procedure used to derive the expression above is known as integration along rays. It only takes into account the constraints (4.1) along rays.[12] The term inside the expectation is

[12] Note that the gradient of the rent function,

$$\nabla r(\theta) = \nabla_\theta c(\theta, y(\theta)),$$

is a conservative vector field since, under assumption A1,

$$\frac{\partial}{\partial\theta_j}\left(\frac{\partial c}{\partial y}(\theta, y(\theta))\right) = \frac{\partial}{\partial\theta_i}\left(\frac{\partial c}{\partial y}(\theta, y(\theta))\right)$$

the virtual surplus. It consists of the first-best social surplus $f(\theta, y(\theta)) - c(\theta, y(\theta))$ plus the distortion needed to prevent deviations along rays $\theta \cdot \nabla_\theta c(\theta, y(\theta)) \frac{q(\theta)}{p(\theta)}$ (which is negative because $c_{\theta_i}(\theta, y) < 0$).

Remark 11 *Armstrong (1996) characterizes the optimal "cost-based" tariff assuming homogeneity of the utility function with respect to the type parameter and separability of the agent's "cost-based" indirect utility. With an additional separability condition on the density function which depends directly on the endogenous separability of the indirect utility, he also shows that this tariff is optimal. McAfee and McMillan (1988) generalize the single-crossing condition for the multidimensional case. However, their condition are so restrictive that imply that the shadow price indifference curves have to be hyperplanes.*

Although we are restricted to utility functions that satisfy Assumption 1, we do not need any homogeneity and separability assumptions. Moreover, the marginal cost indifference curves may not be hyperplanes and, therefore, we do not assume the generalized single-crossing condition. Thus, our setup is not contained neither in Armstrong (1996) nor in McAfee and McMillan (1988).

The following lemma will be useful to characterize the solution of Program (4.13).

Lemma 12 *Suppose the cost function is convex. Let $y : \Theta \to \mathbb{R}$ be a profile of activities and define $\gamma(\theta) \equiv c_y(y(\theta), \theta)$ as the marginal cost associated with it. Let $\tilde{y} : \gamma(\Theta) \to \mathbb{R}$ be a function. Then:*

i. if $y = \tilde{y} \circ \gamma$ and $\tilde{y}$ is decreasing, then $y(\cdot)$ is implementable;

ii. if $y(\cdot)$ is implementable, then there exists a non-increasing $\tilde{y}$ such that $y = \tilde{y} \circ \gamma$.

Proof. First, note that Proposition 26 implies that $y(\cdot)$ is implementable if and only if it satisfies conditions (4.2) and (4.3).

(i) Let us verify conditions (4.2) and (4.3). We have

$$\begin{aligned} c_{\theta_i y}(\theta, y(\theta)) y_{\theta_i}(\theta) &= c_{\theta_i y}(\theta, y(\theta)) \tilde{y}'(\gamma(\theta)) \gamma_{\theta_i}(\theta) \\ &= c_{\theta_i y}(\theta, y(\theta)) \tilde{y}'(\gamma(\theta)) [c_{yy}(\theta, y(\theta)) y_{\theta_i}(\theta) + c_{\theta_i y}(\theta, y(\theta))], \end{aligned}$$

for all $i, j = 1, ..., n$. Therefore, the integration along rays is unimportant to define procedure above.

which implies that

$$[1 - c_{yy}(\theta, y(\theta))\tilde{y}'(\gamma(\theta))]c_{\theta_i y}(\theta, y(\theta))y_{\theta_i}(\theta) = [c_{\theta_i y}(\theta, y(\theta))]^2\tilde{y}'(\gamma(\theta)).$$

Since $c(\theta, y)$ is convex and $\tilde{y}(\cdot)$ is decreasing, it follows that (4.2) holds.

Let $\theta, \hat{\theta} \in \Theta$ be such that $y(\theta) = y(\hat{\theta})$. Since $\tilde{y}(\cdot)$ is decreasing, $\gamma(\theta) = \gamma(\hat{\theta})$. Therefore, $c_y(\theta, y(\theta)) = c_y(\hat{\theta}, y(\hat{\theta}))$ or, equivalently,

$$c_y(\theta, y(\theta)) = c_y(\hat{\theta}, y(\theta)).$$

Hence, (4.3) holds.

(ii) Now, suppose that conditions (4.2) and (4.3) hold. From condition (4.3), it follows that

$$y(\theta) = y(\hat{\theta}) \Rightarrow \gamma(\theta) = \gamma(\hat{\theta}),$$

for all $\theta, \hat{\theta} \in \Theta$. This means that the indifference curve of $y(\cdot)$ are contained in the indifference curves of $\gamma(\cdot)$. Using condition (4.2), we have, through radial directions, that

$$y(\theta) > y(\hat{\theta}) \Rightarrow \gamma(\theta) \leq \gamma(\hat{\theta}),$$

for all $\theta, \hat{\theta}$ in a given radius from 0. Again using (4.3), the last inequality holds in all Θ. Applying the representation theorem for preferences, there must exist a non-increasing $\tilde{y} : \gamma(\Theta) \to \mathbb{R}$ such that $y(\theta) = \tilde{y} \circ \gamma(\theta)$, for all $\theta \in \Theta$. ■

Lemma 12 and Proposition 1 imply that $y(\theta)$ is implementable by $w(y)$ if (4.1) holds and there exists a non-increasing function $\tilde{y}(\gamma)$ such that $y(\theta) = \tilde{y}(\gamma(\theta))$, where $\gamma(\theta) = c_y(\theta, y(\theta))$. Therefore, there is no loss of generality in considering the indirect mechanism where the message space corresponds to the set of possible marginal costs of taking the action y. Each type reveals its marginal cost $\gamma(\theta)$ of taking the prescribed action, takes the action $\tilde{y}(\gamma(\theta))$, and receives a transfer of $w\left(\tilde{y}(\gamma(\theta))\right)$.

Taking the conditional expectation of Program (4.13) and applying the law of iterated

expectations, we obtain the following first-order condition:

$$E\left[f_y(\theta, \tilde{y}(\gamma(\theta)))|\gamma(\theta) = \gamma\right] - \gamma + E\left[\theta.\nabla_\theta c_y(\theta, y(\theta)) \frac{q(\theta)}{p(\theta)}|\gamma(\theta) = \gamma\right] = 0. \tag{4.14}$$

Suppose that there exists an implicit decreasing and non-negative solution of equation (4.14), $\tilde{y}^*(\gamma)$.[13] Then, the following theorem establishes that $y^*(\theta) = \tilde{y}^*(\gamma(\theta))$ is the solution of Program (4.13):

Theorem 3 *Suppose the valuation function f concave in y and the cost function c is convex in y and satisfies Assumption 1. If equation (4.14) defines a decreasing function $\tilde{y}^* : \gamma(\Theta) \to \mathbb{R}_+$ which is integrable,[14] then $y^*(\theta) = \tilde{y}^*(\gamma(\theta))$ is an optimal profile of activities.*

Equation (4.14) has an intuitive interpretation in terms of projections. The unrestricted optimum of Program (4.13) is the pointwise maximization of the virtual surplus. However, this solution may not satisfy conditions (4.2) and (4.3). From Lemma 12, any profile of actions that satisfies these conditions can be written as an indirect profile that is a function of θ only through the marginal cost of taking the action $\gamma(\theta)$. Then, condition (4.15) states that *the solution of the principal's program is determined by the first-order condition of the projection of the virtual surplus on the space of marginal costs $\gamma(\Theta)$.*

Recall that, from Assumption 1, the cost function is $c(\theta, y) = \xi(y) + y \times \psi(\theta) + \varphi(\theta)$. In order to prove Theorem 3, it is useful to consider indirect mechanisms where the message space consists of $\psi(\Theta)$. Although the space of marginal costs $\gamma(\Theta)$ is more intuitive than the space $\psi(\Theta)$, it is harder to work with since the marginal cost $\gamma(\theta)$ is a function of the profile of actions $y(\cdot)$, which is endogenous. However, when costs are convex, working with both message spaces is equivalent:

Lemma 13 *Consider a convex cost function satisfying Assumption 1, let $y : \Theta \to \mathbb{R}$ be a profile of activities, and define $\gamma(\theta) \equiv c_y(\theta, y(\theta))$ as the marginal cost associated with it. There exists a strictly increasing function $\phi : \gamma(\Theta) \to \psi(\Theta)$ such that $\phi(\gamma(\theta)) = \psi(\theta)$ for all $\theta \in \Theta$.*

---

[13] If this relaxed solution does not satisfy the monotonicity condition one has to perform the usual "ironing principle" (see Mussa and Rosen, 1978). Notice that since $y(\gamma)$ is one-dimensional function, this is straightforward exercise.

[14] *Integrability is ensured if, for example, $\lim_{y\to\infty} f(\theta, y) - c(\theta, y) = -\infty$ for all θ. This implies that the implicit solution of (4.14) is bounded. Because Θ is compact, it follows that it is integrable.*

Proof. Since $\gamma(\theta) = \xi'(y(\theta)) + \psi(\theta)$, it follows that

$$\gamma - \xi'(z(\gamma)) = \psi.$$

The left hand side is an increasing function of γ because ξ' is a non-decreasing function (remember that $\xi'' = c_{yy} \geq 0$). Thus, γ is an increasing transformation of ψ, which concludes the proof. ■

Therefore, there is no loss of generality in considering indirect mechanisms where each type sends message $\psi(\theta)$, takes an action $\hat{y}(\psi(\theta))$ and obtains a transfer $w(\hat{y}(\gamma(\theta)))$. Proceeding as in equation (4.14), we obtain:

$$E\left[f_y(\theta, \hat{y}(\psi(\theta)))|\psi(\theta) = \psi\right] - \xi'(y) - \psi(\theta) + E\left[\theta.\nabla_\theta \psi(\theta) \frac{q(\theta)}{p(\theta)}|\psi(\theta) = \psi\right] = 0. \qquad (4.15)$$

Lemma 14 *Suppose the valuation function f concave in y and the cost function c is convex in y and satisfies Assumption 1. If equation (4.15) defines a decreasing function $\hat{y}^* : \psi(\Theta) \to \mathbb{R}$ which is integrable, then $y^*(\theta) = \hat{y}^*(\psi(\theta))$ is an optimal profile of activities.*

Proof. Suppose $\hat{y}(\psi)$ satisfies (4.15) and let $\hat{z}(\psi)$ be an arbitrary implementable profile of activities. By the previous lemma we can suppose that $\hat{y}$ and z are non-increasing functions. Let $y(\theta) := \hat{y}(\psi(\theta))$ and $z(\theta) := \hat{z}(\psi(\theta))$.

Define the following functional

$$F[z] := E\left[D(\theta, z(\theta))\right],$$

where

$$D(\theta, z(\theta)) := f(\theta, z(\theta)) - c(\theta, z(\theta)) + \theta.\nabla c(\theta, z(\theta)) \frac{q(\theta)}{p(\theta)}.$$

Note that $F[z]$ consists of the objective function from Program (4.13) evaluated at $z(\theta)$.

Since $D(\theta, \cdot)$ is a concave function for each θ,

$$D(\theta, z(\theta)) - D(\theta, y(\theta)) \leq D_y(\theta, y(\theta))\left[z(\theta) - y(\theta)\right].$$

Taking the law of iterated expectations, yields

$$F[z] - F[y] \leq E\left\{E\left[D_y(\theta, \hat{y}(\psi))|\psi\left(\theta\right) = \psi\right]\left[\hat{z}(\psi) - \hat{y}(\psi)\right]\right\}.$$

However, (4.15) implies that $E[D_y(\theta, \hat{y}(\psi))|\psi\left(\theta\right) = \psi] = 0$. Thus, $F[z] - F[y] \leq 0$. ■

Theorem 3 then follows immediately from Lemmata 13 and 14. Note that, because $\nabla_\theta y(\theta) = \widehat{y}'(\psi)\nabla_\theta\psi(\theta)$,

$$\widehat{y}'(\psi)E\left[\theta \cdot \nabla_\theta\psi(\theta)\frac{q(\theta)}{p(\theta)}|\psi(\theta) = \psi\right] = E\left[\theta \cdot \nabla_\theta y(\theta)\frac{q(\theta)}{p(\theta)}|\psi(\theta) = \psi\right].$$

Substituting in equation (4.15), we obtain

$$\begin{aligned} & \{E[f_y(\widehat{y}(\psi), \theta)|\psi(\theta) = \psi] - \gamma\} \times E\left[\theta \cdot \nabla_\theta y(\theta)\frac{q(\theta)}{p(\theta)}|\psi(\theta) = \psi\right] \qquad (4.16) \\ = \; & -\widehat{y}'(\psi)\left(E\left[\theta \cdot \nabla_\theta\psi(\theta)\frac{q(\theta)}{p(\theta)}|\psi(\theta) = \psi\right]\right)^2 \geq 0, \end{aligned}$$

where the inequality uses the fact that, by Lemmata 12 and 13, $\widehat{y}$ is non-increasing. Note that equation (4.1) implies that $\gamma\left(\theta\right) = w'\left(y\left(\theta\right)\right)$. Therefore, the inequality above generalizes condition (4.9) for the multidimensional case.

The following examples illustrate the usefulness of the characterization in Theorem (3):

Example 14 (One-dimensional model with SCC) *Take Θ to be an interval in $\mathbb{R}$ and let $c_{\theta y} < 0$, $c_{yy} \geq 0$. Since $n = 1$, we have*

$$\theta q(\theta) = \int_1^\infty \theta p(\tau\theta)d\tau = 1 - P\left(\theta\right).$$

Because $\gamma(\theta)$ is a decreasing function, we can apply a change of variables from γ to θ. Then, equation (4.14) becomes

$$f_y\left(\theta, y\left(\theta\right)\right) - c_y\left(\theta, y\left(\theta\right)\right) = -c_{\theta y}\left(\theta, y\left(\theta\right)\right)\frac{1 - P\left(\theta\right)}{p\left(\theta\right)},$$

which is the standard first-order condition of the unidimensional relaxed problem.

Example 15 (One-dimensional Labor Market without SCC) *In the model from Example 9, let* $\beta = 0$.[15] *The cost function becomes* $\hat{c}(\theta_1, y, \bar{g}) = \frac{y}{\theta_1(\bar{g}-\alpha\theta_1)}$, *which satisfies Assumption 1. The marginal cost of schooling is* $\gamma(\theta_1) = \psi(\theta_1) = \frac{1}{\theta_1(\bar{g}-\alpha\theta_1)}$. *Suppose that* θ_1 *conditional on* $\bar{g}$ *is uniformly distributed on* $\left[\underline{\theta}, \bar{\theta}\right]$. *In the appendix, we show that condition (4.14) becomes:*

$$f_y\left(\frac{\bar{g} - \sqrt{\bar{g}^2 - \frac{4\alpha}{\psi}}}{2\alpha}, \hat{y}(\psi)\right) + f_y\left(\frac{\bar{g} + \sqrt{\bar{g}^2 - \frac{4\alpha}{\psi}}}{2\alpha}, \hat{y}(\psi)\right) = \bar{g}\left(\bar{\theta} + \frac{\bar{g}}{\alpha}\right)\psi^2 - 2\psi. \tag{4.17}$$

This equation characterizes the solution of the signaling model presented by Araujo, Gottlieb, and Moreira (2007) in a screening environment. If f *is type-independent (so that it can be written as* $f(y)$*), the solution is* $\hat{y}(\psi) = f_y^{-1}\left(\frac{\bar{g}}{2}\left(\bar{\theta} + \frac{\bar{g}}{\alpha}\right)\psi^2 - \psi\right)$.

Example 16 (Two-dimensional Labor Market) *Let* $\theta = (\theta_1, \theta_2)$ *be uniformly distributed on* $[0,1]^2$ *and suppose the valuation function is type-independent (so that it can be written as* $f(y)$*). As in Example 9, take* $c(\theta, y) = \frac{y}{\theta_1\theta_2}$. *The marginal cost function is given by* $\gamma(\theta) = \psi(\theta) = \frac{1}{\theta_1\theta_2}$. *Then, condition (4.14) becomes*

$$f'(\widetilde{y}(\gamma)) = \gamma\left\{1 + E\left[\frac{q(\theta)}{p(\theta)}\middle|\gamma(\theta) = \gamma\right]\right\}.$$

In the appendix, we show that $E\left[\frac{q(\theta)}{p(\theta)}\middle|\gamma(\theta) = \gamma\right] = \frac{1}{2\gamma} - \frac{1}{2\gamma^2} - \frac{\ln(\gamma)}{2\gamma^2}$. *Thus, the solution is characterized by*[16]

$$f'(\tilde{y}(\gamma)) = \frac{2\gamma^2 + \gamma - 1 - \ln(\gamma)}{2\gamma},$$

or, in terms of the agent's type,

$$f'(y(\theta_1, \theta_2)) = \frac{1}{\theta_1\theta_2} + \frac{1}{2} - \frac{\theta_1\theta_2}{2}\left[1 - \ln(\theta_1\theta_2)\right].$$

Example 17 (Nonlinear Pricing) *Consider the model of Example 11. Since* $f_y = MC$ *and*

[15] *See Araujo and Moreira (2004) for the case* $\beta \neq 0$. *The analysis of this case is more complex since implementability is not necessary by the local and global conditions presented in this paper.*

[16] *It is straightforward to show that the implicit solution* $\tilde{y}(\gamma)$ *is decreasing in* γ.

$\gamma(\theta) = c_y(\theta, y(\theta)) = V_Q\left(\theta, \bar{Q} - y(\theta)\right)$, *equation (4.14) becomes*

$$V_Q(\theta, Q(\theta)) = MC + E\left[\theta.\nabla_\theta V_Q(\theta, Q(\theta))\frac{q(\theta)}{p(\theta)}|\gamma(\theta) = \gamma\right].$$

When types are one-dimensional and the SCC is satisfied, this equation reduces to the standard nonlinear pricing rule:

$$V_Q(\theta, Q(\theta)) = MC + V_{\theta Q}(\theta, Q(\theta))\frac{1 - F(\theta)}{f(\theta)}.$$

The example below relates condition (4.16) to its one-dimensional counterpart (4.9):

Example 18 (Multidimensional Screening) *Take $n \geq 2$ and suppose that the distribution function is homogeneous of degree $\alpha < 2 - n$. By homogeneity, $\frac{q(\theta)}{p(\theta)} > 0$ can be factored out of the conditional expectation in condition (4.16). Thus, we obtain*

$$\left\{E[f_y(\theta y(\theta))|\psi(\theta) = \psi] - w'(y(\theta))\right\} E[\theta \cdot \nabla_\theta y(\theta)|\psi(\theta) = \psi] \geq 0,$$

which is the multidimensional equivalent of condition (4.9).

Suppose that the valuation function is type-independent (i.e., it can be written as $f(y)$) and the cost cost function is homogenous of degree $\beta < 0$ on θ.[17] *Then, the first-order condition (4.14) yields*

$$f'(\widetilde{y}(\gamma)) - \gamma + a\beta\gamma = 0,$$

where $a = \int_1^\infty \tau^{\alpha+n-1}d\tau < \infty$ since $\alpha < 2 - n$. From (1), this condition, after multiplying by $\widetilde{y}'(\gamma)$, becomes

$$\left[f'(\widetilde{y}(\gamma)) - w'(\widetilde{y}(\gamma))\right]\widetilde{y}'(\gamma) = a\beta\gamma\widetilde{y}'(\gamma) \leq 0.$$

Thus, in this case, condition (4.9) holds when we identify each type by its marginal cost of the activity γ (as was shown in the proof of Lemma 12, $\tilde{y}$ is decreasing).

Reciprocally, for each mechanism $(y(\cdot), w(y))$ we are able to find an economy such that the mechanism is the solution of the principal's program. This is formally stated in the following

[17] *Armstrong (1996) assumes $\beta = -1$.*

theorem:

Theorem 4 *Suppose* $\Theta = \mathbb{R}^n_+$. *Let* $y : \Theta \to \mathbb{R}_+$ *be a regular function and* $w(y)$ *be a positive* C^2 *function. There exist a valuation function linear in* y, *a* C^1 *cost function satisfying Assumptions 0 and 1, and a distribution of types* p *for which* $(y(\theta), w(y))$ *is the optimal mechanism if and only if* $w(y)$ *is strictly increasing and condition (4.16) is satisfied.*

Proof. ($\Rightarrow$) Follows from an adaptation of Theorem 1.

($\Leftarrow$) Let $(y(\theta), w(y))$ be a mechanism satisfying the conditions of the Theorem. Applying Theorem 1, we can find a cost function satisfying Assumptions A0 and A1 such that this mechanism is incentive compatible. Notice that $\gamma(\theta) = w'(y(\theta))$. Take a density p over Θ that is homogeneous of degree $\alpha < 2-n$. Hence, maximizing (4.13) pointwise is equivalent to setting $f(\theta, y) = \mu(\gamma)y$, where

$$\mu(\gamma) = \gamma - aE\left[\theta.\nabla_\theta c_y(\theta, y(\theta))|\gamma(\theta) = \gamma\right]$$

satisfies equation (4.15), which is necessary and sufficient for optimality, and a is defined in the previous example. ■

The participation constraint (IR) implies that all types participate in the mechanism. When $\varphi(\theta) + \xi(0) = 0$ for all θ (as in Examples 14, 15, 16) there is no loss of generality in assuming so. In general, however, it may be optimal to exclude some types. In that case, it is useful to distinguish between two types of models:

1. Certain Participation: $\varphi(\theta) = \bar{\varphi}$ for all $\theta \in \Theta$,
2. Random Participation: $\varphi(\theta)$ is non-constant in $\theta \in \Theta$.

Next, we study the exclusion of types in models with certain and random participation separately.

Certain Participation

Without loss of generality, we can normalize $\bar{\varphi} = 0$. It is straightforward to show that the exclusion region is defined by $\Theta_0 = \{\theta \in \Theta; \gamma(\theta) \geq \gamma_0\}$, where γ_0 is such that $\tilde{y}(\gamma_0) = 0$.

Thus, it follows that, if equation (4.14) defines a decreasing function $\tilde{y}^* : \gamma(\Theta) \to \mathbb{R}$ which is integrable, the optimal profile of activities is given by

$$y^*(\theta) = \max\{\tilde{y}^*(\gamma(\theta)), 0\}.$$

Equivalently, if equation (4.15) defines a decreasing function $\hat{y}^* : \psi(\Theta) \to \mathbb{R}$, then the optimal profile of activities is $y^*(\theta) = \max\{\tilde{y}^*(\psi(\theta)), 0\}$.

Random Participation

Recall that, by Lemmata 12 and 13, there is no loss of generality in considering indirect mechanisms where the message space is $\psi(\Theta)$. Let $\hat{r}(\psi) = w(\hat{y}(\psi)) - \xi(\hat{y}(\psi)) - \psi\hat{y}(\psi)$. Given any fixed mechanism, types which participate are those whose outside options are lower than $\hat{r}$, i.e.,

$$\hat{r}(\psi) \geq \varphi(\theta).$$

Then, the principal's expected payoff is

$$E\left\{[f(\theta, \hat{y}(\psi)) - \xi(\hat{y}(\psi)) - \psi\hat{y}(\psi) - \hat{r}(\psi)]\,\mathbf{1}_{[\hat{r}(\psi)\geq\varphi(\theta)]}\right\},$$

where $\mathbf{1}$ denotes the indicator function. If the solution $\hat{y}(\psi)$ of this program is a decreasing function, then $y(\theta) = \hat{y}(\psi(\theta))$ is the solution of the principal's program.

Assume that the valuation function f is type-independent. Hence, as in models of second-degree price discrimination, the agent's parameter of private information does not enter the principal's payoff directly. Then, applying the law of iterated expectations, the principal's payoff is

$$E\left\{[f(\hat{y}(\psi)) - \xi(\hat{y}(\psi)) - \psi\hat{y}(\psi) - \hat{r}(\psi)]\,E\left[\mathbf{1}_{[\hat{r}(\psi)\geq\varphi(\theta)]}\middle|\,\psi(\theta) = \psi\right]\right\}.$$

Let $G^{\varphi}(\psi, x) = E\left[\mathbf{1}_{[x\geq\varphi(\theta)]}\middle|\,\psi\right]$ denote the probability that a type with rent $r(\theta) = x$ participates conditional on $\psi(\theta) = \psi$. Denote the social surplus by $S(\psi, y) \equiv f(y) - \xi(y) - \psi y$. Following the approach of Rochet and Stole (2002), the principal's program can be written as

the maximization of

$$E\left\{\left[S\left(\psi,\hat{y}\left(\psi\right)\right)-\hat{r}\left(\psi\right)\right]G^{\varphi}\left(\psi,\hat{r}\left(\psi\right)\right)\right\},$$

subject to $\hat{y}\left(.\right)$ being non-increasing and there existing a $w\left(.\right)$ such that $w(\hat{y}(\psi))=\hat{r}(\psi)+\xi(\hat{y}(\psi))+\psi\hat{y}(\psi)$.

By the envelope theorem, this program is equivalent to

$$\max_{\hat{y}(.)} E\left\{\left[S\left(\psi,\hat{y}\left(\psi\right)\right)-\hat{r}\left(\psi\right)\right]G^{\varphi}\left(\psi,\hat{r}\left(\psi\right)\right)\right\}$$

$$\text{subject to } \hat{r}'\left(\psi\right)=-\hat{y}\left(\psi\right), \text{ and}$$

$$\hat{y}\left(\psi\right) \text{ non-increasing.}$$

Ignoring the monotonicity condition, the solution is characterized by the second-order differential (Euler equation):

$$\frac{d}{d\psi}\left[G^{\varphi}\left(\psi,\hat{r}\left(\psi\right)\right)S_{y}\left(\hat{y}\left(\psi\right),\psi\right)\right]+\frac{\partial}{\partial r}\left\{\left[S\left(\psi,\hat{y}\left(\psi\right)\right)-\hat{r}\left(\psi\right)\right]G^{\varphi}\left(\psi,\hat{r}\left(\psi\right)\right)\right\}=0,$$

with boundary conditions $\hat{y}\left(\psi_{m}\right)=y^{FB}(\theta_{m})$ and $\hat{y}\left(\psi_{M}\right)=y^{FB}(\theta_{M})$, where $\psi_{m}=\min\psi\left(\theta\right)$, $\theta_{m}=\arg\min_{\theta\in\Theta}\psi\left(\theta\right)$, $\psi_{M}=\max\psi\left(\theta\right)$, $\theta_{M}=\arg\max_{\theta\in\Theta}\psi\left(\theta\right)$, $y^{FB}\left(\theta\right)$ is the first-best solution (i.e., it satisfies $S_{y}(\theta,y^{FB}(\theta))=0$), and $\hat{r}'(\psi)=-\hat{y}(\psi)$.

Simplifying and using the fact that $\hat{r}'(\psi)=-\hat{y}(\psi)$, we obtain:

$$G^{\varphi}\left(\psi,\hat{r}\left(\psi\right)\right)\left[\left(\xi''\left(\hat{y}\left(\psi\right)\right)-f''\left(\hat{y}\left(\psi\right)\right)\right)\hat{r}''\left(\psi\right)-2\right]+\frac{\partial}{\partial\psi}G^{\varphi}\left(\psi,\hat{r}\left(\psi\right)\right)\left[f'\left(\hat{y}\left(\psi\right)\right)-\xi'\left(\hat{y}\left(\psi\right)\right)-\psi\right] \tag{4.18}$$

$$+\frac{\partial}{\partial r}G^{\varphi}\left(\psi,\hat{r}\left(\psi\right)\right)\left[\left(f'\left(\hat{y}\left(\psi\right)\right)-\xi'\left(\hat{y}\left(\psi\right)\right)\right)\hat{r}\left(\psi\right)+f\left(\hat{y}\left(\psi\right)\right)-\xi\left(\hat{y}\left(\psi\right)\right)-\hat{r}\left(\psi\right)\right]=0.$$

Assuming that equation (4.18) implicitly defines a non-increasing function $\hat{y}\left(\psi\right)$, it characterizes the solution of the random participation model when valuations are type-independent. If the solution $\hat{y}\left(\psi\right)$ is not non-increasing, the solution is obtained by applying an ironing procedure. The proposition below summarizes this result:

Proposition 27 *Suppose the valuation function f is type-independent and concave and suppose*

that the cost function c is convex in y and satisfies Assumption 1. If equation (4.18) defines a decreasing function $\hat{y}^* : \psi(\Theta) \to \mathbb{R}$ which is integrable, then $y^*(\theta) = \hat{y}^*(\psi(\theta))$ is an optimal profile of activities.

Equation (4.18) generalizes the characterization of Rochet and Stole (2002) for arbitrary distributions of types and arbitrary cost functions satisfying Assumption 1. The following example shows that their model can be obtained as a special case of our characterization:

Example 19 (Rochet and Stole, 2002) *Let $\Theta \subset \mathbb{R}^2$ be an interval and denote types by $(t, x) \in \Theta$. Let $\psi(t, x) = -t$, $\varphi(t, x) = -x$, $f(y) - \xi(y) = \frac{y^2}{2}$. Assume that types t and x are independently distributed and denote by $f(t)$ and $G(x)$ the probability distribution of t and the cumulative distribution of x, respectively. Let $M(t, u) \equiv G^{\varphi}(t, u) = f(t)G(u)$.*

Then, equation (4.18) becomes

$$M_u(t,u)\left(u - \frac{1}{2}\dot{u}^2\right) + M(t,u)(2 - \ddot{u}) + M_t(t,u)(t - \dot{u}) = 0,$$

which is the equation obtained by Rochet and Stole.

Therefore, this section analyzed the solution of screening models under Assumption 1. Theorem 3 characterized the solution when all types participate. In 4.3.2, it was shown that this characterization can be easily generalized to the case of certain participation (i.e., $\varphi(\theta) = \bar{\varphi}$ for all θ). In 4.3.2, we characterized the solution with the exclusion of types when participation is random but valuation functions are type-independent. This characterization generalized the one presented by Rochet and Stole (2002).

In terms of empirical implications, Theorem 4 established that the screening model imposes two restrictions on the space of mechanisms $(y(\theta), w(y))$. First, by incentive-compatibility, it requires w to be monotonic. Second, maximization of the principal's payoff imposes condition (4.16). When types are one-dimensional or when the distribution function and the cost function are homogeneous and the valuation function is type-independent, this condition implies that the principal's profit as a function of types must increase at a greater rate under asymmetric information than the increase in productivity (which is equal to the rate of growth under symmetric information).

4.4 The Signaling Game

In this section, we consider a standard signaling game where preferences are as described in Section 4.2. There are many identical principals ('receivers') who act competitively. For simplicity, we consider the case where the principal's valuation function does not depend on the amount of the activity ('signal') exerted by the agent ('sender'):

$$f(\theta, y) = f(\theta).$$

Therefore, the only way through which the activity affects the transfer is through its informational content. Of course, our results extend to the case where the activity also affects the valuation. Our competitive assumption implies that the transfer is equal to the expected valuation of the sender conditional on the signaling activity.

The timing of the signaling game is as follows. First, nature determines the type of each sender, θ, according to the density function p. Then, senders choose the amount of signaling y contingent on their types. Subsequently, the market offers a transfer $w(y)$ conditional on the observed signal.

Since all receivers are equal, we will study symmetric equilibria where the offered wage schedule is the same for every receiver. As usual, we adopt the perfect Bayesian equilibrium concept. In what follows $E[\cdot|\cdot]$ represents the conditional expectation operator with respect to the measure of beliefs μ.

Definition 9 *A perfect Bayesian equilibrium (PBE) for the signaling game is a profile of strategies* $(y(\theta), w(y))$ *and beliefs* $\mu(\cdot \mid y)$ *such that:*

1. *The sender's strategy is optimal given the equilibrium wage schedule, i.e., (IC) holds.*

2. *The market is competitive (i.e., receivers earn zero profits):*

$$w(y) = E[f(\theta) \mid y(\cdot) = y]. \tag{4.19}$$

3. *Beliefs are consistent:* $\mu(\theta \mid y)$ *is derived from the sender's strategy using Bayes' rule where possible.*

Substituting the zero-profit condition (4.19) into the first-order condition from incentive-compatibility (4.1), we obtain:

$$y_{\theta_i}(\theta) = \frac{\frac{\partial}{\partial \theta_i} E\left[f(\cdot)|y(\cdot) = y(\theta)\right]}{c_y\left(\theta, y\left(\theta\right)\right)}. \tag{4.20}$$

Consider a separating equilibrium. Bayes' rule implies that $w\left(y\left(\theta\right)\right) = E\left[f(\cdot)|y(\cdot) = y(\theta)\right] = f\left(\theta\right)$ for all θ. Therefore, equation (4.20) becomes $y_{\theta_i}\left(\theta\right) = \frac{f_{\theta_i}(\theta)}{c_y(\theta, y(\theta))}$. Therefore, Assumption 0 implies that we must have $y_{\theta_i}\left(\theta\right) > 0$ in any separating equilibrium. However, from the second-order condition from incentive compatibility (4.2), we can only have $y_{\theta_i}\left(\theta\right) > 0$ if $c_{\theta_i y}\left(\theta, y\left(\theta\right)\right) \leq 0$. Hence, *a fully separating equilibrium exists only if* $c_{\theta_i y}\left(\theta, y\left(\theta\right)\right) \leq 0$ *for all* θ*: The SCC is satisfied along the equilibrium signal* y.[18]

Given an equilibrium profile $(y(\cdot), w(\cdot))$, denote the type with the lowest amount of activity y, the amount he chooses, and the wage he gets by

$$\begin{aligned} \theta^{\min} &= \arg\min_{\hat{\theta} \in \Theta} y\left(\hat{\theta}\right), \\ y^{\min} &= y(\theta^{\min}), \text{and} \\ w^{\min} &= w\left(\theta^{\min}\right). \end{aligned} \tag{4.21}$$

Let $f^{\min}$ denote the lowest valuation in the economy:

$$f^{\min} = \min_{\hat{\theta} \in \Theta} f\left(\hat{\theta}\right).$$

In a PBE we need to make sure that agents have no incentive to deviate to actions off the equilibrium path. In order to obtain a sufficient condition for types not to benefit from deviating to actions off the equilibrium path, let beliefs off the equilibrium be given by

$$\mu\left(\theta|y\right) = \begin{cases} 1 \text{ if } \theta = \arg\min f\left(\hat{\theta}\right) \\ 0 \text{ if } \theta \neq \arg\min f\left(\hat{\theta}\right) \end{cases},$$

for $y \notin y\left(\Theta\right)$. By deviating to any $y \notin y\left(\Theta\right)$, an agent gets transfer $f^{\min}$, which gives a payoff of

[18] This condition is slightly less demanding than the SCC, which states that $c_{\theta_i y}\left(\theta, y\right) < 0$ for all θ, y.

at most $f^{\min} - c(\theta, 0)$. Then, type θ does not want to deviate to any action off the equilibrium path $y \notin y(\Theta)$ if this payoff is lower than $E\left[f(\cdot) \mid y^{\min}\right] - c\left(\theta, y^{\min}\right)$. Therefore, types do not benefit from deviating to actions off the equilibrium path if $y^{\min} \geq 0$ satisfies

$$E\left[f(\theta) \mid y^{\min}\right] - f^{\min} \geq c\left(\theta, y^{\min}\right) - c(\theta, y), \quad \forall \theta, \tag{4.22}$$

for all $y \notin y(\Theta)$.

Notice that the set of actions $y^{\min} \geq 0$ such that this inequality is satisfied is non-empty (since $y^{\min} = 0$ satisfies this condition). Moreover, when $\theta^{\min}$ is separated, the only $y^{\min}$ compatible with (4.22) is $y^{\min} = 0$. Inequality (4.22) gives boundary conditions for the differential equation (4.1).

Conditions (4.1), (4.2), (4.3), (4.19), and (4.22) are necessary for a PBE. As in Example 13, they may not be sufficient when Assumption 1 does not hold. However, the following lemma states that they are sufficient for a PBE when Assumption 1 holds.

Lemma 15 *Suppose Assumption 1 holds and let $y(\cdot)$ and $w(\cdot)$ be regular functions. There exists a set of beliefs $\mu(\cdot \mid y)$ such that $(y(\cdot), w(\cdot), \mu(\cdot))$ is a PBE if and only if the first- and second-order conditions (4.1) and (4.2), the pooling condition (4.3), the zero-profit condition (4.19), and the boundary condition (4.22) are satisfied.*

Proof. ($\Rightarrow$) Straight from Proposition 26.

($\Leftarrow$) Define the transfer schedule as in Proposition 26, where $w^{\min}$ is as defined in (4.21).

Observe that $w(y^{\min}) = w^{\min}$ and, by (4.20), $w'(y(\theta))y_{\theta_i}(\theta) = \frac{\partial}{\partial \theta_i} E[f(\cdot)|y(\cdot) = y(\theta)]$ for every regular type θ and every i. Therefore, by continuity, $w(y) = E[f(\theta) \mid y(\cdot) = y]$ for all $y \in y(\Theta)$, i.e., the zero profit condition holds.

Lemma 26 has shown that the agent's strategy is optimal given the equilibrium transfer schedule. This concludes the proof. ■

Inequality (4.22) implies that, when $\theta^{\min}$ is not separated, there may exist several boundary conditions $y^{\min}$ that satisfy the requirements of a PBE. In order to deal with the multiplicity of equilibria, we will follow part of the literature by selecting based on an efficiency criterion.

Thus, we impose as a selection criterion that

$$y^{\min} = 0. \tag{4.23}$$

We will refer to a PBE satisfying condition (4.23) as a *least costly equilibrium.* The proposition below states necessary and sufficient conditions for a least costly equilibrium.

Proposition 28 *Suppose Assumption 1 holds and let $y(\cdot)$ and $w(\cdot)$ be regular functions. There exists a set of beliefs $\mu(\cdot \mid y)$ such that $(y(\cdot), w(\cdot), \mu(\cdot))$ is a least costly equilibrium if and only if the first- and second-order conditions (4.1) and (4.2), the pooling condition (4.3), and the boundary condition (4.23) are satisfied.*

Proof. Follows straight from Lemma 15. ■

Proposition 28 characterizes the least costly equilibria of the signaling model. This characterization allows us to study which restrictions follow from a signaling model when the single-crossing condition is not imposed.

We know from Section 4.2 that incentive-compatibility implies that the transfer schedule must be strictly increasing. Furthermore, our selection criterion implies that there must be some type θ such that $y(\theta) = 0$. The theorem below shows that these are the only implications of the signaling model.

Theorem 5 *Let $y(\cdot)$ be a regular function and let $w(\cdot)$ be a positive C^2 function. There exists a C^1 cost function satisfying Assumption 1 and a distribution of types p for which $(y(\cdot), w(\cdot))$ is a least costly equilibrium profile of signals and transfers if and only if $w(\cdot)$ is strictly increasing and there exists a $\theta \in \Theta$ such that $y(\theta) = 0$. Furthermore, this equilibrium profile is the same for all distribution of types p.*

Proof. ($\Rightarrow$) Obvious.

($\Leftarrow$) Let us define the following C^1 functions: $c(\theta, y)$ as in the proof of Theorem 1 and

$$f(\theta) = w(y(\theta)).$$

Observe that c and f are non-negative C^1 functions. We claim that the pair $(y(\cdot), w(\cdot))$ are the profiles of signals and transfers in a least costly equilibrium for the economy $\{c, f, p\}$, for any

density p. First, by Proposition 26 $(y(\cdot), w(\cdot))$ is incentive compatible which implies Condition 1 of the definition of the PBE. Moreover, $y(\theta) = y(\widetilde{\theta}) \Rightarrow f(\theta) = f(\widetilde{\theta})$, which implies that $E[f(\cdot)|y(\theta) = y] = f(\theta)$ and $\frac{\partial}{\partial \theta_i}\{E[f(\cdot)|y(\theta) = y]\} = f_{\theta_i}(\theta)$, for all θ and density p. We then only need to prove the first-order conditions (4.20) which is equivalent to

$$\frac{f_{\theta_i}(\theta)}{c_y(\theta, y(\theta))} = \frac{w'(y(\theta))y_{\theta_i}(\theta)}{w'(y(\theta))} = y_{\theta_i}(\theta).$$

Finally, the boundary condition (4.23) obviously holds. Using Theorem 1, we conclude the proof. ■

Since the equilibrium constructed in Theorem 5 holds for all distributions p and for any cost function $c(\theta, y) = w'(y(\theta))y + \frac{\hat{K}}{2}(y - y(\theta))^2$ with $\hat{K} > K$, it follows that the model is not identified given data on signals and transfers. Indeed, any distribution of types is compatible with the same equilibrium profile of signals and transfers. Thus, we have the following corollary:

Corollary 6 *Signaling models are not identified given data on signals and transfers: for every profile $(y(\cdot), w(\cdot))$ satisfying the conditions of Theorem 5, there is an infinite number of signaling models that have this profile as a least costly equilibrium.*

Remark 12 *The only robust property that emerges from the equilibrium is the monotonicity of the wage schedule. If the cost of signaling function is increasing on the signal then, by revealed preference, equilibrium wages are increasing.*

Therefore, in the context of education as a signal, the only robust implication of the signaling hypothesis is the monotonicity of wages in education. However, this result is also shared by the usual human capital (symmetric information) models. Indeed, the revealed preference argument holds regardless of the link between education and productivity.

In the context of advertising as a signal, Theorem 5 implies that revenues must be increasing in advertisement. Because this is implied by revealed preference it holds regardless of the relationship between quality and advertising. It is also shared by Becker and Murphy's (1993) model of advertisement as a good, for example.

Remark 13 *For one-dimensional type models, Theorem 5 says that signaling models are compatible with non-monotonic signaling functions. However, from (4.2), this is only possible*

when the SCC is violated. Recent works show that such non-monotonic equilibria may emerge and have important empirical consequences [see Araujo et al. (2007) and Araujo and Moreira (2003)].

Theorem 5 does not allow one to control for the valuation function. From an applied perspective, it assumes that the econometrists do not observe the valuation function. Next, we show that even if one controls for the valuation function, the indeterminacy result still remains. More precisely, let $y(\theta)$ be a profile of signals and $w(y)$ be the associated transfer schedule that satisfy the conditions of Theorem 2. Fix any valuation function $f(\theta)$. We say that f is consistent with the given profile of signals and transfers if it satisfies the zero-profit condition:

$$w(y) = E[f(\cdot)|y(\theta) = y], \tag{4.24}$$

where the expectation is taken with respect to some distribution of types p. Definition 9 implies that this consistency requirement must be satisfied in any PBE. The following corollary establishes that any profile of signals and transfers satisfying (4.24) can be rationalized by a cost of signaling function and a distribution of types as a part of a least costly equilibrium of this economy (characterized by the fixed valuation function and the distribution of types). Formally, we have the following:

Corollary 7 *Let $y(\cdot)$ be a regular function, let $w(\cdot)$ be a C^2 function, and let $f(\cdot)$ be a valuation function satisfying (4.24) with respect to a continuous density p. There exists a C^1 cost function satisfying Assumption 1 for which $(y(\cdot), w(\cdot))$ is a least costly equilibrium profile of signals and transfers if and only if $w(\cdot)$ is strictly increasing and there exists a $\theta \in \Theta$ such that $y(\theta) = 0$.*

Proof. ($\Rightarrow$) Obvious.

($\Leftarrow$) Take the cost of signaling exactly as in the proof of Theorem 1 and the density p. Then, following the same steps of the proof of Theorem 5 establishes the result. ■

Theorem 5 and Corollary 7 show that there are only three implications of signaling models. First, the zero-profit condition implies that transfers have to be in the convex hull of the valuation of agents. Second, our selection criterion based on efficiency requires some type to choose zero amount of the activity (least costly equilibrium). However, this result is not

robust since there are PBEs that do not satisfy this condition. Third, the fact that signaling is costly implies that transfers must be strictly increasing in signals. Although this conclusion is extremely robust, it is also implied by most alternative models. Any profiles of activities y and transfers w satisfying these three conditions is compatible with a least costly equilibrium of a signaling model.

Remark 14 (The multidimensional case $n \times n$) *By Quinzii and Rochet (1985) another interesting extension to consider here is economies that have the same dimension of types and signals. Do the result of our paper remain the same? Under a natural generalization of assumption A0 and A1 we have the following conclusions: (i) the characterization of incentive compatibility Section 2 are naturally generalizable; (ii) the indeterminacy results of this section (Theorem 5 and Corollary 1) are then straightforward; (iii) however, the characterization of the screening problem is not trivial. Although one can extend the integration by rays to get an analogous expression of Program (13), it is not obvious that Lemma 3 will be true in this case and consequently the projection method to derive an analogous first-order condition (14).*

4.5 Conclusion

This paper studied incentive-compatibility when the single-crossing condition is not satisfied. This allowed us to provide a characterization of the solution of multidimensional screening models and the equilibria of multidimensional signaling models. Then, using our characterization, we analyzed the restrictions imposed by incentive-compatibility, screening, and signaling once the single-crossing condition does not hold.

First, it was shown that the only implication of incentive-compatibility was the monotonicity of transfers in actions. Then, in the case of screening models, we obtained a new necessary and sufficient condition. In the one-dimensional models, it implies the principal's profit to grow at a higher rate under asymmetric information than the it would grow under symmetric information.

In the case of signaling models, the zero-profit condition implies that transfers must be in the convex hull of valuations in each set of pooled types. We have also imposed a selection criterion that requires some type to choose a zero amount of the signal. We have shown that any profile of actions and transfers satisfying these conditions is an equilibrium of some economy.

Therefore, apart from these mild restrictions, the implications of signaling and screening models arise from assumptions of the cross-partial derivative of the cost function. In the absence of more precise knowledge of the cost function, we cannot obtain other testable predictions from these models.

Appendix A One-dimensional Labor Market Model without SCC

First note that $\psi = \frac{1}{\theta_1(\overline{g}-\alpha\theta_1)}$ implies that

$$\theta_1 = \frac{\overline{g} - \sqrt{\overline{g}^2 - \frac{4\alpha}{\psi}}}{2\alpha} \text{ or } \theta_1 = \frac{\overline{g} + \sqrt{\overline{g}^2 - \frac{4\alpha}{\psi}}}{2\alpha}.$$

In a discrete pooling set these types are equally likely conditional to $\psi(\theta) = \psi$ under the uniform distribution. We have that

$$\frac{\theta_1 q(\theta_1)}{p(\theta_1)} = \frac{1 - P(\theta_1)}{p(\theta_1)} = \overline{\theta} - \theta_1$$

and, since $\psi'(\theta_1) = -(\overline{g} - 2\alpha\theta_1)\psi^2 = -\frac{\psi}{\theta_1} + \alpha\theta_1\psi^2$,

$$\begin{aligned} \frac{\theta_1 q(\theta_1)}{p(\theta_1)}\psi'(\theta_1) &= (\overline{\theta} - \theta_1)\left(-\frac{\psi}{\theta_1} + \alpha\theta_1\psi^2\right) \\ &= \overline{\theta}\left(-\frac{\psi}{\theta_1} + \alpha\theta_1\psi^2\right) + \psi - \alpha\theta_1^2\psi^2. \end{aligned}$$

Therefore,

$$E\left[\frac{\theta_1 q(\theta_1)}{p(\theta_1)}\psi'(\theta_1)\middle|\, \psi(\theta_1) = \psi\right] = -\overline{\theta}\psi E\left[\frac{1}{\theta_1}\middle|\, \psi(\theta_1) = \psi\right] + \psi - \alpha\psi^2 E\left[\theta_1^2\middle|\, \psi(\theta_1) = \psi\right]$$

but

$$\begin{aligned} E\left[\frac{1}{\theta_1}\middle|\, \psi(\theta_1) = \psi\right] &= \frac{\psi\overline{g}}{2} \text{ and} \\ E\left[\theta_1^2\middle|\, \psi(\theta_1) = \psi\right] &= \frac{\overline{g}^2}{2\alpha^2} - \frac{1}{\alpha\psi}. \end{aligned}$$

This implies that

$$E\left[\frac{\theta_1 q(\theta_1)}{p(\theta_1)}\psi'(\theta_1)\middle|\psi(\theta_1)=\psi\right] = -\frac{\overline{\theta}\overline{g}}{2}\psi^2+\psi-\frac{\overline{g}^2}{2\alpha}\psi^2+\psi$$
$$= 2\psi-\frac{\overline{g}}{2}\left(\overline{\theta}+\frac{\overline{g}}{\alpha}\right)\psi^2.$$

Finally, condition (4.14) becomes

$$f_y\left(\frac{\overline{g}-\sqrt{\overline{g}^2-\frac{4\alpha}{\psi}}}{2\alpha},\hat{y}(\psi)\right)+f_y\left(\frac{\overline{g}+\sqrt{\overline{g}^2-\frac{4\alpha}{\psi}}}{2\alpha},\hat{y}(\psi)\right)=\overline{g}\left(\overline{\theta}+\frac{\overline{g}}{\alpha}\right)\psi^2-2\psi,$$

which is the condition (4.17).

Appendix B Two-dimensional Labor Market Model

Let $p=1$ on $[0,1]^2$ and $\gamma=\frac{1}{\theta_1\theta_2}$. Thus, $\theta\cdot\nabla_\theta\gamma=-\gamma$, and

$$q(\theta)=\int_1^\infty \tau^{n-1}p(\tau\theta)d\tau=\left[\int_1^{\min\{1/\theta_1,1/\theta_2\}}\tau d\tau\right]^+=\frac{1}{2}\left[\min\{1/\theta_1,1/\theta_2\}^2-1\right]^+.$$

Lemma: For all $\gamma\in[1,\infty)$,

$$E\left[\frac{q(\theta)}{p(\theta)}\middle|\gamma(\theta)=\gamma\right]=\frac{1}{2\gamma}-\frac{1}{2\gamma^2}+\frac{\ln(\gamma)}{2\gamma^2}.$$

Proof: We have to show that

$$E\left[E\left[\frac{q(\theta)}{p(\theta)}\middle|\gamma(\theta)=\gamma\right]h(\gamma)\right]=E\left[\frac{q(\theta)}{p(\theta)}h(\gamma)\right]$$

for mensurable function $h:[1,\infty)\to\mathbb{R}$. Let us calculate the right hand side. To do it, we will make the change of variable from (θ_1,θ_2) to (θ_1,γ):

$$T:[0,1]^2\to[0,1]\times[1,\infty]$$
$$(\theta_1,\theta_2)\to(\theta_1,1/\theta_1\theta_2)$$

Then,

$$T^{-1} : D \to [0,1]^2$$
$$(\theta_1, \gamma) \to (\theta_1, 1/\gamma\theta_1)$$

where $D = T\left([0,1]^2\right) = \{(\theta_1, \gamma) \in [0,1] \times [1,\infty]; \gamma \geq 1/\theta_1\}$ and

$$|DT^{-1}| = \begin{vmatrix} 1 & 0 \\ -\frac{1}{\gamma\theta_1^2} & -\frac{1}{\gamma^2\theta_1} \end{vmatrix} = -\frac{1}{\gamma^2\theta_1}.$$

Hence,

$$\begin{aligned} E\left[\frac{q(\theta)}{p(\theta)}h(\gamma)\right] &= \int_0^1\int_0^1 q(\theta)h(\gamma)d\theta_2 d\theta_1 \\ &= \int_0^1\int_{1/\theta_1}^1 \widetilde{q}(\theta_1,\gamma)h(\gamma) \times \left(\frac{1}{\gamma^2\theta_1}\right) d\gamma d\theta_1 \\ &= \int_1^\infty\int_{1/\gamma}^1 \widetilde{q}(\theta_1,\gamma)\frac{1}{\gamma^2\theta_1}h(\gamma)d\theta_1 d\gamma \end{aligned} \tag{4.25}$$

where the last equality comes from Fubini's theorem and $\widetilde{q}(\theta_1,\gamma) = q(\theta_1, 1/\gamma\theta_1)$. Note that

$$\begin{aligned} \frac{1}{\theta_1} &\geq \frac{1}{\theta_2} = \gamma\theta_1 \Leftrightarrow \theta_1 \leq \frac{1}{\gamma^{1/2}}, \\ \theta_2 &= \frac{1}{\gamma\theta_1} \leq 1 \Leftrightarrow \frac{1}{\gamma} \leq \theta_1. \end{aligned}$$

Therefore, we have that

$$\widetilde{q}(\theta_1,\gamma) = \frac{1}{2}\begin{cases} \gamma^2\theta_1^2 - 1, & \frac{1}{\gamma} \leq \theta_1 \leq \frac{1}{\gamma^{1/2}} \\ \frac{1}{\theta_1^2} - 1, & \frac{1}{\gamma^{1/2}} \leq \theta_1 \leq 1 \end{cases}.$$

Substituting in equation (4.25), yields

$$E\left[\frac{q(\theta)}{p(\theta)}h(\gamma)\right] = \int_1^\infty \frac{h(\gamma)}{2\gamma^2}\left\{\int_{1/\gamma}^{1/\gamma^{1/2}}\left(\gamma^2\theta_1 - \frac{1}{\theta_1}\right)d\theta_1 + \int_{1/\gamma^{1/2}}^1\left(\frac{1}{\theta_1^3} - \frac{1}{\theta_1}\right)d\theta_1\right\}d\gamma. \tag{4.26}$$

Note that

$$\int_{1/\gamma}^{1/\gamma^{1/2}}\left(\gamma^2\theta_1 - \frac{1}{\theta_1}\right)d\theta_1 = \gamma^2\frac{\theta_1^2}{2}\Big|_{1/\gamma}^{1/\gamma^{1/2}} - \ln(\theta_1)|_{1/\gamma}^{1/\gamma^{1/2}} = \frac{\gamma^2}{2}\left(\frac{1}{\gamma} - \frac{1}{\gamma^2}\right) - \frac{1}{2}\ln(\gamma),$$

and

$$\begin{aligned}\int_{1/\gamma^{1/2}}^{1}\left(\frac{1}{\theta_1^3}-\frac{1}{\theta_1}\right)d\theta_1 &= -\frac{1}{2\theta_1^2}\Big|_{1/\gamma^{1/2}}^{1} - \ln(\theta_1)|_{1/\gamma^{1/2}}^{1} \\ &= -\frac{1}{2}+\frac{\gamma}{2}-\frac{1}{2}\ln(\gamma).\end{aligned}$$

Therefore, equation (4.26) simplifies to

$$E\left[\frac{q(\theta)}{p(\theta)}h(\gamma)\right] = \int_1^{\infty}\frac{h(\gamma)}{2\gamma^2}\left(\gamma-1-\ln(\gamma)\right)d\gamma$$

for all h, which establishes the result.

References

Acemoglu, Daron, 1998. Credit Market Imperfections and the Separation of Ownership from Control. Journal of Economic Theory 78, 355-381.

Araujo, A., Moreira, H., 2003. Non-monotone insurance contracts and their empirical consequences. Discussion Paper #512, Getulio Vargas Foundation.

Araujo, A., Moreira, H., 2004. Adverse Selection Problems without the Spence-Mirrlees Condition. Discussion Paper #425, Getulio Vargas Foundation.

Araujo A., Gottlieb D., Moreira H., 2007. A Model of Mixed Signals with applications to Countersignaling, RAND Journal of Economics 38, 1020-1043.

Araujo, A., Moreira, H., Tsuchida, M., 2004. Do Dividends Signal More Earnings? A theoretical analysis. Discussion Paper #524, Getulio Vargas Foundation.

Araujo, A., Moreira, H., 2003. Non-monotone insurance contracts and their empirical consequences. Discussion Paper #512, Getulio Vargas Foundation.

Armstrong, M., 1996. Multiproduct Nonlinear Pricing, Econometrica 64, 51-75.

Bagwell, L. S., Bernheim, B. D., 1996. Veblen Effects in a Theory of Conspicuous Consumption. American Economic Review 86, 394-377.

Becker, G. S., Murphy, K. M., 1993. A Simple Theory of Advertising as a Good or Bad. Quarterly Journal of Economics 108, 941-964.

Bernheim, B. D., 1991. Tax Policy and the Dividend Puzzle. Rand Journal of Economics 22; 455-476.

Bernheim, B. D., 1994. A Theory of Conformity. Journal of Political Economy 102; 841-877.

Bernheim, B. D., Redding, L. S., 2001. Optimal Money Burning: Theory and Application to Corporate Dividend Policy, Journal of Economics & Management Strategy, 10, 463-507.

Bernheim, B.D., Severinov, S., 2003. Bequests as Signals: An Explanation for the Equal Division Puzzle, Journal of Political Economy, 111, 733-764.

Chiappori, P.-A., Jullien, B., Salanié, B., Salanié, F., 2006. Asymmetric Information in Insurance: General Testable Implications, RAND Journal of Economics, 37, 783-798.

Engers, M., Fernandez, L., 1987. Market Equilibrium with Hidden Knowledge and Self-Selection. Econometrica 55, 425-439.

Heckman, J., 2005. Lessons from the Technology of Skill Formation. NBER Working Paper #11142.

Heckman, J., J. Stixrud, Urzua S., 2005. The Effects of Cognitive and Noncognitive Abilities on Labor Market Outcomes and Social Behavior. Discussion Paper, University of Chicago.

Kohllepel, L. Multidimensional 'Market Signaling', 1983. Discussion Paper, Universität Bonn.

Matthews, S., Moore, J. 1987. Monopoly Provision of Quality and Warranties: An Exploration in the Theory of Multidimensional Screening. Econometrica 55,441-467.

McAfee, R.P., McMillan, J., 1988. Multidimensional incentive compatibility and mechanism design. Journal of Economic Theory 46, 335-354

Milgrom, P. R., Segal, I. R., 2002. Envelope Theorems for Arbitrary Choice Sets, Econometrica 70, 583-601.

Milgrom, P. R., Shannon, C., 1992. Monotone Comparative Statics, Econometrica 62, 157-180.

Mirrlees, J. A., 1971. An Exploration in the Theory of Optimum Income Taxation. Review of Economic Studies 38, 175-208.

Quinzii, M., Rochet, J.-C., 1985. Multidimensional Signaling. Journal of Mathematical Economics 14, 261-284.

Rochet, J.-C., Stole, L., 2003. The Economics of Multidimensional Screening. In: M. Dewatripont, L. P. Hansen and S. J. Turnovsky (eds.) Advances in Economics and Econometrics: Theory and Applications - Eight World Congress, 2003.

Rotemberg. J. J., 1988. Pareto Improving Distortionary Taxes in the Presence of Signaling. Alfred P. Sloan School of Management Working Paper No. 2039-88.

Rothschild. M. and Stiglitz, J. E., 1976. Equilibrium in competitive insurance markets: An essay on the economics of imperfect information. Quarterly Journal of Economics 90, 629-650.

Smart, M., 2000. Competitive Insurance Markets with Two Unobservables. International Economic Review 41, 153-169.

Spence, M., 1974. Market Signalling. Harvard University Press, Cambridge, MA.